AD-A116 424   AMERICAN COLL TESTING PROGRAM IOWA CITY IA   F/G 14/2
SOME STATISTICAL PROCEDURES FOR DOMAIN-REFERENCED TESTING: A HA--ETC(U)
FEB 81   R L BRENNAN   N00123-78-C-1206
UNCLASSIFIED   NPRDC-TN-81-6   NL

1 of 2
AD A
116424

6

NPRDC TN 81-6

FEBRUARY 1981

SOME STATISTICAL PROCEDURES FOR DOMAIN-REFERENCED TESTING: A HANDBOOK FOR PRACTITIONERS

NAVY PERSONNEL RESEARCH
AND
DEVELOPMENT CENTER
San Diego, California 92152

NPRDC Technical Note 81-6

February 1981

SOME STATISTICAL PROCEDURES FOR DOMAIN-REFERENCED TESTING: A HANDBOOK FOR PRACTITIONERS

Robert L. Brennan

American College Testing Program
Iowa City, Iowa

Reviewed by
John D. Ford, Jr.

Approved by
Richard C. Sorenson

FL-182 on file

A

Navy Personnel Research and Development Center
San Diego, California 92152

FOREWORD

This effort was conducted under Contract N00123-78-C-1206 with the American College Testing (ACT) Program within work unit ZF522.012.03.01, Criterion-Referenced Testing (CRT). The objective of this work unit is to develop and evaluate innovative CRT techniques to alleviate some of the deficiencies and problems that exist with current procedures used in the Navy training/testing community (e.g., item-writing methods, item statistics, generalizing to the domain of performance, and computerized adaptive testing).

The purpose of the ACT effort was to investigate errors of measurement in criterion-referenced, domain-referenced, and mastery testing. This effort has been conducted in two phases. NPRDC TN 80-15 (Brennan, 1980a) reported on the first phase: The development of a computer program to estimate error variances, variance components, and indices of dependability. That technical note tells testing researchers how to run the program, and how to interpret and use the results appropriately.

This technical note reports on the second phase: The development of a handbook of some simple statistical techniques for producing and evaluating criterion and/or domain-referenced test (DRTs) for Navy technical training. It is a "how-to-do-it" handbook for use in developing and assessing CRTs and/or DRTs. Specifically, it considers item analysis procedures, techniques for establishing cutting scores, errors of measurement and classification, test length, and advancement scores, as well as group-based coefficients of agreement.

This handbook is a working document intended for limited distribution to Center personnel and peers in the scientific community. It is not a formal presentation of Center research. Parts of it will be incorporated into a larger, more comprehensive testing manual for achievement and diagnostic testing that is being produced by NAVPERSRANDCEN for the Navy technical training community.

The contracting officer's technical representative was Pat-Anthony Federico.

RICHARD C. SORENSON
Director of Programs

SUMMARY

Problem

Many of the statistical techniques that have been used for developing and evaluating norm-referenced tests are not applicable to criterion-referenced tests (CRTs) and domain-referenced tests (DRTs) since the data from these later tests do not usually follow the normal distribution. Further, CRTs and DRTs are not used to compare or rank students against one another; rather, they are used to determine whether students have met or exceeded mastery learning levels or absolute performance standards. Statistical procedures are needed that can be easily employed by developers and evaluators of CRTs and DRTs in the Navy.

Purpose

The purpose of this effort was to investigate errors of measurement in criterion-referenced, domain-referenced, and mastery testing.

Approach

A handbook of some statistical techniques for producing and evaluating DRTs was created for Navy practitioners. This is a "how-to-do-it" guide for the intelligent layman who develops and assesses DRTs and/or CRTs. This handbook considers item analysis procedures, techniques for establishing cutting scores, errors of measurement and classification, test length, and advancement scores, as well as group-based coefficients of agreement.

Results and Conclusions

No attempt was made to catalogue, list, or describe exhaustively a large number of available procedures for a particular purpose. Rather, a few procedures were selected for a single purpose based upon the principal investigator's judgment as to which are the best techniques for that purpose. Simple numerical examples were used to illustrate procedures, and guidelines were provided for using and/or interpreting results.

Future Direction

This handbook will be incorporated into a larger, more comprehensive testing manual for achievement and diagnostic testing, which is being produced by NAVPERSRANDCEN for the Navy technical training and testing community.

Table of Contents

List of Tables

List of Tables (Continued)

List of Figures

1. Introduction

Almost twenty years ago, the term "criterion-referenced testing" was introduced into the literature on educational measurement, and since that time an enormous number of papers have been published that deal with technical issues in this area. In no way does this handbook represent an attempt to synthesize all of this literature; rather, this handbook treats a restricted set of statistical procedures for addressing some of the most prevalent technical issues that arise in criterion-referenced testing, which is frequently called "domain-referenced testing."

Throughout this handbook the term "domain-referenced" will be used instead of "criterion-referenced" for two principal reasons. First, the term "criterion-referenced" too readily suggests some external criterion against which examinee performance on a test can be compared. There are situations in which an external criterion exists and relevant data are available. However, such situations are rare in this author's experience; and, indeed, none of the procedures discussed in this handbook require criterion data, in the usual sense of the word "criterion." Second, in this handbook it is assumed that the items in a test can be viewed as a sample from a larger universe of potential items that might have been chosen for the test. It is natural to refer to this universe as a domain--hence, the term "domain-referenced."

In domain-referenced testing, the examinee's score of principal interest is the examinee's score over *all* items in the *universe* of items. This score can never be obtained directly, but it can be estimated by, for example, the examinee's observed score on a set of items, or test. Also, in domain-referenced testing, the interpretation of an examinee's score is *not* based on the scores obtained by other examinees. In a sense, therefore, the phrase "domain-referenced testing" is itself something of a misnomer, because what is of principal interest is domain-referenced *interpretations* of examinee scores. Such interpretations are frequently contrasted with norm-referenced interpretations that involve comparing the performance of an examinee *relative* to the performance of other examinees.

To put it another *way*, even highly qualified experts would have great difficulty distinguishing between a norm-referenced and a domain-referenced test, per se; and the procedures for administering and scoring domain-referenced and norm-referenced tests seldom differ much at all. What *is* different is the interpretations given to the resulting scores, and the procedures employed to study the quality of these interpretations. Indeed, in principle, scores on any test can be given either domain-referenced or norm-referenced interpretations, although this is rarely done.

Actually, one can distinguish between two types of domain-referenced interpretations. One interpretation rests on using an examinee's observed score on a test as an estimate of his/her universe score. The other interpretation involves comparing an examinee's score to some fixed cutting score that is defined independently of examineee test scores. This latter type of interpretation is frequently associated with mastery/non-mastery decisions.

The procedures discussed in this handbook do not necessarily represent the most technically sophistocated procedures available. Indeed, the procedures discussed here have been chosen, in large part, because they do not necessitate extensive computations, even though the theoretical foundation for some of these procedures is highly technical. Also, no claim is made that the procedures discussed in this handbook treat all relevant issues in domain-referenced testing, although they do cover those issues most frequently discussed. The general intent is simply to provide practitioners with a unified treatment of some relatively straightforward statistical procedures for use in domain-referenced testing.

Sample Statistics

In this handbook all computational formulas and procedures are provided in tables that include examples employing synthetic data. In every case, the computations involve nothing more mathematically complicated than computing sample means, variances, and standard deviations , and then combining these quantities in various ways.

It is assumed here that the reader is already at least partially familiar with the concepts of a mean, variance, and standard deviation. In a certain statistical sense, a mean is a single number (an average value, or a "central" value) that represents an entire set of scores, while variance and standard deviation are convenient measures of the amount of spread, or dispersion, in a set of scores.

Table 1.1 provides formulas for calculating the sample statistics used in this handbook. To give the statistics in Table 1.1 a concrete interpretation, the formulas for them are expressed with respect to a person's mean score, or _proportion_ of items correct (number of items answered correctly divided by the total number of items). In this handbook, a person's mean score is represented $\bar{x}_p$, where the "bar" over the variable x signifies a mean score, and the subscript _p_ signifies a particular person. Specifically, if there are _n_ items and x_{pi} represents the score for the person _p_ on item _i_, then the mean score for person _p_ is $\bar{x}_p = \Sigma_i x_{pi}/n$, where Σ_i means "the sum over items."

If one wanted to express these sample statistics in terms of a person's total score on a test, then the symbol x_p (without a bar) would be used. Also, it should be noted that what is important is the _form_ of the equations in Table 1.1--not the fact that they are expressed in terms of a variable x. The same form would apply if the variable were labelled y, as is the case in one section of this handbook.

In Table 1.1 two formulas are provided for sample variance--one that uses the symbol s^2, and one that used the symbol $\hat{s}^2$. The latter is easily obtained from the former, and in almost all cases these two statistics will have very similar (but _not_ identical) values. In certain sections of this handbook, s^2 is used, and in other sections $\hat{s}^2$ is used.

Table 1.1

Formulas *for* Calculating Sample Means, Variances, and Standard Deviations

Formulas	Example
Let $\bar{x}_p$ = observed mean score for person p (proportion of items correct)	Suppose k = 6 persons have the following observed mean scores: .6, .8, .8, .8, .9, .9
k = number of persons	
Σ = a symbol meaning "sum the scores"	
Calculate	
$\Sigma\, \bar{x}_p$ = sum of mean scores for k persons	$\Sigma\, \bar{x}_p = .6 + .8 + .8 + .8 + .9 + .9 = 4.80$
$\Sigma\, \bar{x}_p^2$ = sum of squared mean scores for k persons	$\Sigma\, \bar{x}_p^2 = .36 + .64 + .64 + .64 + .81 + .81 = 3.90$
Sample Mean:	
(1.1) $\bar{x} = \Sigma\, \bar{x}_p/k$	$\bar{x} = 4.80/6 = .80$
Sample Variance:	
(1.2) $s^2(\bar{x}_p) = \dfrac{\Sigma\, \bar{x}_p^2}{k} - \bar{x}^2$	$s^2(\bar{x}_p) = \dfrac{3.90}{6} - (.80)^2 = .010$
Sample Standard Deviation	
(1.3) $s(\bar{x}_p) = \sqrt{s^2(\bar{x}_p)}$	$s(\bar{x}_p) = \sqrt{.010} = .100$
Corrected Sample Variance	
(1.4) $\hat{s}^2(\bar{x}_p) = \dfrac{k}{k-1}\, s^2(\bar{x}_p)$	$\hat{s}^2(\bar{x}_p) = \dfrac{6}{6-1}\,(.010) = .012$
Corrected Sample Standard Deviation	
(1.5) $\hat{s}(\bar{x}_p) = \sqrt{\hat{s}^2(\bar{x}_p)}$	$\hat{s}(\bar{x}_p) = \sqrt{.012} = .110$

However, as far as this handbook is concerned, the sole reason for choosing between s^2 and $\hat{s}^2$ is to provide the simplest possible computational procedures for estimating quantities of interest. (A similar statement holds for the corresponding standard deviations, s and $\hat{s}$.)

It was mentioned, above, that a standard deviation is a measure of the amount of spread or dispersion in a set of scores. To give the concept of a standard deviation a more concrete interpretation, it is common practice to consider the standard deviation of a particular bell-shaped distribution of scores, called a normal distribution. As illustrated in Figure 1.1, for a normal distribution: (a) 68% of the scores lie within one standard deviation to the right and left of the mean; and (b) 95% of the scores lie within two standard deviations to the right and left of the mean. These two statements also can be expressed in terms of what are called "z-scores."

As indicated in Figure 1.1, a score that lies one standard deviation above the mean can be denoted $z = 1$; and, a score that lies one standard deviation below the mean can be denoted $z = -1$. It follows that, for a normal distribution, 68% of the scores lie between $z = -1$ and $z = 1$. Similarly, 95% of the scores lie between $z = -2$ and $z = 2$.

The above statements about percent of cases between specified z-scores do not apply to all possible distributions of scores. However, provided one does not interpret such statements too literally, they can properly serve as useful bench marks for conceptualizing the interpretation of a standard deviation.

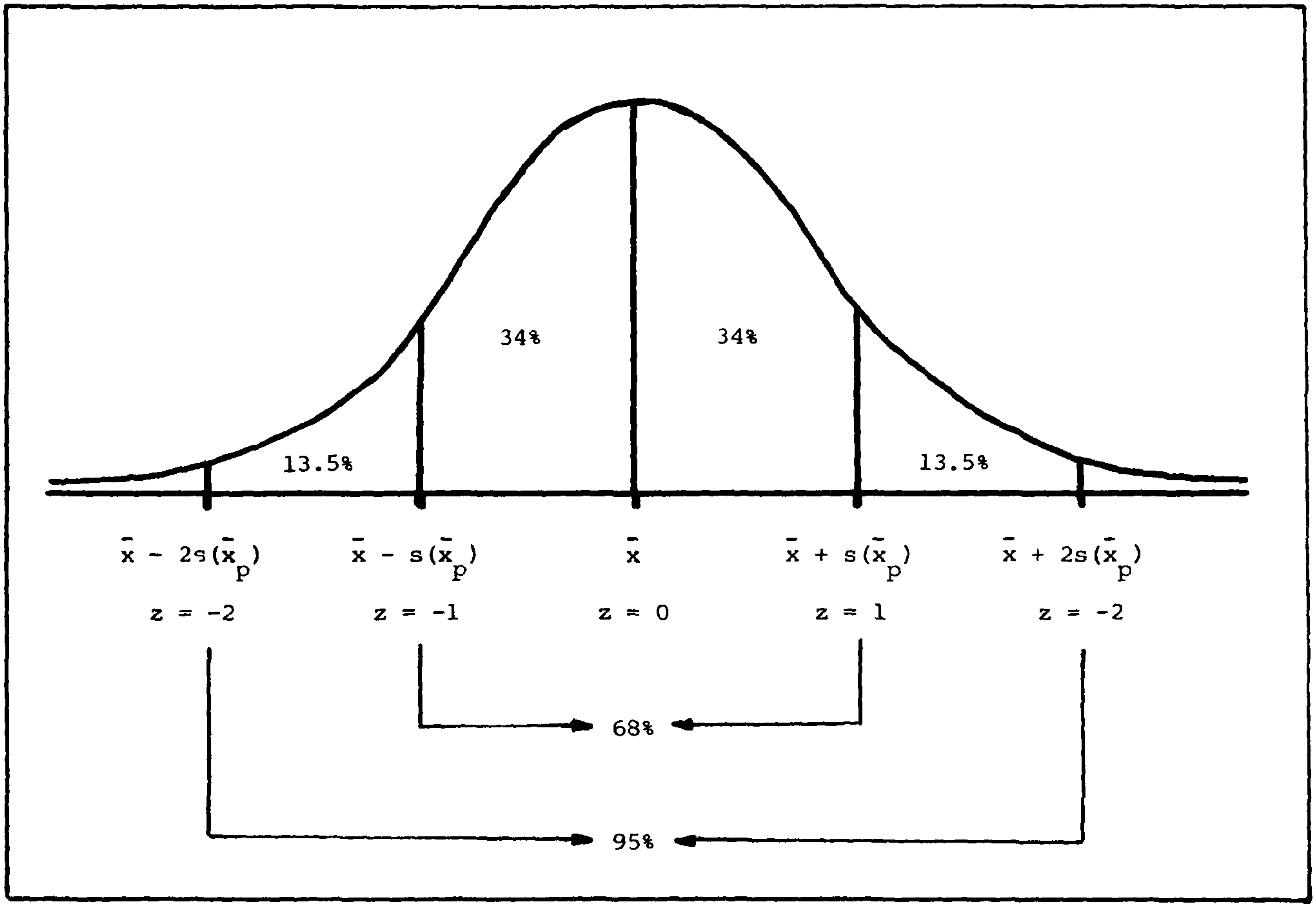

Figure 1.1. Normal Distribution

The reader is cautioned not to infer from the above paragraphs that test scores are usually (or should be) normally distributed. Indeed, for domain-referenced tests, it is quite common to have many high-scoring examinees and relatively few low-scoring examinees; and such a distribution is not normal. For this reason, most procedures treated in this handbook involve no assumption about the shape of the score distribution.

Universe of Items

A universe of items is a concept of central importance for domain-referenced interpretations, because ultimately one wants to make inferences about examinee universe, or domain, scores. (Considerations with respect to a universe of items are prominent in some approaches to norm-referenced interpretations, too, but norm-referenced interpretations are not within the scope of this handbook.)

Sometimes there actually exists a set of items that can be considered as the intended universe. For example, some computer-managed instruction systems have a large bank of items that is used to construct specific tests. Also, the words in a specified dictionary might constitute a universe for a spelling domain.

More frequently, however, pragmatic concerns require that one conceptualize a universe of items for the content under consideration. For example, in the initial stages of developing a domain-referenced testing system, it is likely that only a limited number of items will be available. Furthermore, for many content areas, it would be virtually impossible to construct all relevant items, or even a large proportion of such items.

In such cases, it is especially important that the intended universe be defined and described in as clear and unambiguous a manner as possible. Otherwise, one cannot easily claim that a particular item does, or does not, reference the intended domain; nor can one clearly specify what an examinee's universe score means.

No matter how a universe may be defined, in this handbook a test is viewed as a sample of items from an intended universe. More specifically, to be technically correct, we ought to say that a test is a random sample of items from the universe, in the sense that every item in the universe has an equal chance of appearing in any test. In practice, one seldom has the opportunity to randomly select a sample of items, in the literal sense of the word "randomly." However, if a universe is defined well enough, then one can usually ensure that a test consists of a reasonably representative sample of items from the intended universe.

It can be argued that for every objective in a program or instructional sequence, there ought to be a distinct universe of items. It is not uncommon, however, for a test to reference a universe that might be viewed as stratified, in the sense that the universe is defined by multiple objectives or the multiple categories in a table of specifications or task-content matrix. The procedures discussed in this handbook do not specifically incorporate considerations with respect to a universe defined in this manner, even though these procedures (or similar ones) are sometimes used with such universes.

Overview

No matter how well-defined a universe of items may be, the quality of the decisions made can be no higher than the quality of the items themselves. Therefore, Section 2 considers some simple item analysis procedures for using data to help identify items that may be flawed. This topic is rather mundane, and the process of performing item analyses is tedious; but, in this author's opinion the validity of a domain-referenced measurement procedure absolutely necessitates using good items that represent a well-defined universe of items. Furthermore, no after-the-fact statistical analysis of examinee test scores can overcome the negative impact of poor items on the quality of domain-referenced interpretations.

Section 3 considers a rather simple procedure for establishing a cutting score, π_o , expressed as a proportion of items correct for the universe of items. (In this handbook the Greek letter π is used to represent a score for the universe of items, whereas x is used for a score on a test, or sample of items from the universe.) This procedure is "content-based" in the sense that it relies upon the subjective (but, hopefully, well-informed) judgments of content-matter specialists.

Section 4 treats a procedure for establishing an advancement score. Recall that a cutting score, π_o , is expressed as a proportion of items correct for the universe of items; and, as such, π_o is "similar" to an examinee's universe score, π, in the sense that both π_o and π reference the same universe of items. By contrast, an advancement score, x_o , is "similar" to an examinee's observed score, x, in the sense that both reference a test score. To put it another way, an advancement score is an observed score analogue of a cutting score, just as an examinee's test score is an observed score analogue of his/her universe score. A decision concerning mastery is actually *made* with respect to the advancement score; i.e., an examinee is declared a master if his/her observed score is at or above the advancement score.

Section 5 considers two types of error that can be made when a decision about an examinee is based on the examinee's observed score rather than his/her universe score (which is never known). These two types of error are called error of measurement and error of classification. Error of measurement involves the extent to which examinee observed and universe scores differ; and, as such, error of measurement does not involve consideration of a cutting score. By contrast, an error of classification is made if an examinee is erroneously classified as a master or erroneously classified as a non-master.

Section 6 considers a number of issues associated with assessing the quality of domain-referenced measurement procedures for a *group* of examinees. These issues are, in part, related to traditional notions of reliability (or measurement consistency). Also, to an extent, these issues have a validity connotation, because in domain-referenced testing, examinee universe scores are a principal "criterion" of interest. However, the terms "reliability" and "validity" are used only infrequently in Section 6 because they too easily connote traditional statistical analyses (for norm-referenced interpretations) that are inappropriate in domain-referenced measurement contexts. Rather, emphasis is placed upon certain agreement coefficients and group-based measures of error.

Restrictions in Scope and Content

Domain-referenced measurement is currently a topic of considerable interest in numerous applied settings, and a handbook such as this cannot treat all relevant issues in all such settings. In particular, there are many important educational, philosophical, legal, ethical, and technical issues involved in testing for licensure, certification, "minimal" competency, etc. For the most part, such issues are not treated here; rather emphasis is placed upon procedures that seem to this author to be both theoretically reasonable and capable of being used relatively easily by practitioners--especially practitioners in instructional and training environments where nothing more sophisticated than a simple hand-held calculator may be available.

Throughout this handbook it is assumed that examinee responses are not corrected for guessing. In several cases, the procedures discussed could be (or have been) modified in various ways to take guessing into account. Such modifications are not treated here for three reasons. First, many such modifications make assumptions about guessing that the author believes are unrealistic. Second, reasonable assumptions about guessing involve complexities considerably beyond the scope of this handbook. Third, it remains to be seen (in a research sense) whether or not procedures involving reasonable assumptions about guessing materially improve the quality of decisions made in typical domain-referenced testing situations.

In the field of statistics, distinctions are carefully drawn between quantities of principal interest, called parameters, and estimates of these quantities, called statistics. For theoretical work, this distinction is crucial, but to incorporate this distinction in the body of this handbook would necessitate a much more complicated notational system, as well as considerably more complex verbal statements. Therefore, the term "statistic" is used in this handbook in a generic sense (even though occasionally the word "parameter" would be better, technically), and there is no notational distinction drawn between parameters and estimates. Also, both quantities of principal interest and their estimates are usually denoted with Greek letters to distinguish them from the sample statistics discussed in conjunction with Table 1.1. Finally, concerning notational conventions, sometimes a symbol is underlined in the text for emphasis and/or to preclude mistaking it for part of a word or phrase.

The body of this handbook does not contain references to published work, proofs of formulas and equations, or justifications for choosing the procedures treated here rather than others which might have been chosen. However, to a limited extent, these issues are treated in Appendix B, which is provided principally for the technically oriented reader. It will be evident to such a reader that, in several cases, the treatments of procedures in the body of the handbook are slight modifications of procedures discussed in published literature. Such modifications were made principally for computational convenience. Furthermore, in a few instances procedures are presented, or suggestions are made, that have not been considered previously in published literature.

2. Item Analysis Considerations

In domain-referenced testing (or any type of testing, for that matter) there is no substitute for good items. No statistical procedure can overcome the negative effect of poor test items; but as discussed in this section, statistics can be used to help identify poor items.

First, however, it must be emphasized that, prior to collecting any data, every effort must be made to insure that items reflect the objectives they are intended to measure and that the items have no obvious technical flaws. Such judgments are best made by content matter specialists who have knowledge of item construction procedures and guidelines. If content-matter specialists do not have such knowledge then they should be aided in their judgments by someone who does. Also, items should be reviewed for potential bias by members of minority groups, especially when domain-referenced tests are to be used with members of minority groups.

Item Analysis Table and Statistics

No matter how thoroughly content matter experts scrutinize items to eliminate flaws, it is always advisable to study examinee responses to items. Such data provide an additional check on item quality. Usually such data are displayed in the form of an item analysis table such as that provided in Table 2.1.

To give a context to the synthetic data in Table 2.1, let us assume that 10 items were administered to 50 examinees, and one of these items

Table 2.1

Illustration of an
Item Analysis Table and Statistics
Using Synthetic Data

	Subgroup[a]					
Alternative	Low (0-6)	Medium (7-8)	High (9-10)	Total	p	B
a	3	1	2	6	.14	-.13
b*	8	9	16	33	.75	.18
c	2	1	1	4	.09	-.10
d	0	0	0	0	.00	.00
Omit	0	0	1	1	.02	.05
Not Reached	3	3	0	6	--	--
Total	16	14	20	50		
Total minus Not Reached	13	11	20	44	--	--

$$(2.1) \quad p = \left[\begin{array}{l}\text{proportion of examinees who choose}\\ \text{alternative (or omitted item)}\end{array}\right]$$

$$(2.2) \quad B = \left[\begin{array}{l}\text{proportion of examinees}\\ \text{in high group who choose}\\ \text{alternative (or omitted}\\ \text{item)}\end{array}\right] - \left[\begin{array}{l}\text{proportion of examinees}\\ \text{in low group who choose}\\ \text{alternative (or omitted}\\ \text{item)}\end{array}\right]$$

e.g. For the correct alternative, b,

$p = 33/44 = .75$

$B = (16/20) - (8/13) = .80 - .62 = .18$

[a] Numbers within parentheses indicate the scores (in terms of number of items correct) that fall into each group.

Note. * indicates the correct (keyed) alternative.

resulted in the data in Table 2.1. Table 2.1 indicates that this item contains four alternatives with the correct (or keyed) alternative being b (the alternative that is starred). Note that the other alternatives (namely a, c, and d) are sometimes called distractors, or incorrect alternatives.

To study examinee performance on an item, it is usual to classify the examinees into groups based on their test performance. In Table 2.1 this has been accomplished by assigning each examinee to: (a) a "low" group if he/she has 0 - 6 items correct; (b) a "medium" group if he/she has 7 - 8 items correct; or (c) a "high" group if he/she has 9 - 10 items correct. For present purposes, the reader can assume that examinees in the high group would be judged "successful," those in the low group would be judged "unsuccessful," and those in the middle group might (or might not) be judged "successful."

The entries under the columns headed low, medium, and high are the numbers of examinees in each group who chose each alternative, omitted the item, or did not reach the item. The following procedure can be used to distinguish between an item that was omitted (but attempted) by an examinee and one that was not reached (and unattempted): (a) if an examinee omitted the last item, assume that the examinee did not reach one item; (b) if the examinee omitted both of the last two items assume that two items were not reached by the examinee; (c) if the examinee omitted all three of the last three items, assume that three items were not reached; etc. All other blank responses by an examinee can be treated as "omits."

Table 2.1 also includes column totals indicating the total number of examinees in each group, and the number of examinees in each group who reached the item. The row totals in Table 2.1 indicate the total number of examinees who picked each alternative, omitted the item, or did not reach the item. Finally, for each alternative, Table 2.1 provides two statistics which are identified as p and B and defined in Equations 2.1 and 2.2, respectively. The statistic p will always have a value between 0 and 1, and B will always be between -1 and +1.

The statistic p indicates the proportion of examinees who chose an alternative. For the correct alternative p is called the *item difficulty level*, and it is the proportion of examinees who got the item correct. In Table 2.1, $p = .75$ for the correct alternative. Note that *easy* items have *high* difficulty levels and *hard* items have *low* difficulty levels.

The statistic B indicates the difference between the proportions of examinees in the high and low groups who chose an alternative. For the correct alternative, B is called an *item discrimination index*. It reflects the difference between the proportion of examinees in the high group who got the item correct and the proportion in the low group who got the item correct.

Using Item Analysis Data

The principal use of item analysis data in domain-referenced testing situations is to detect flawed items. It must be understood, however, that such data--no matter how carefully analyzed--do not provide an absolute indication that an item is or is not flawed. Also, if an item *is* flawed, the data cannot tell the investigator exactly how to correct the flaw. What the data can do is flag a potentially flawed item and usually suggest the nature of the problem and/or the part of the item that is flawed. Given this perspective, the following paragraphs provide some guidelines for examining item analysis data.

(a) Have an actual copy of the item available when examining an item analysis table like that in Table 2.1.

(b) Look at *p* for the correct alternative. The item may be flawed if the item difficulty level, *p*, is considerably out of line with a value one might expect. (Usually, in domain-referenced testing items have relatively high difficulty levels if they are obtained for a group of examinees who have experienced instruction in the content tested.)

(c) Look at the relationship between item difficulty level and the *p* values for the distractors. If a distractor has a value for *p* that is *above* the item difficulty level, then, examine the distractor to see if in fact it could be considered, reasonably, as a correct answer. If so, one of three problems probably exist--the correct answer was mis-specified, the item has two or more correct answers, or the item is ambiguous. In any case, the item requires revision.

(d) If *p* is very small for any distractor (e.g., alternative *d* in Table 2.1) *consider* eliminating it or replacing it with some other incorrect alternative--*provided* doing so does not change the intended nature of the item. (Recall that if an item is inherently easy, it is very likely that one or more distractors will be chosen infrequently.)

(e) Look at the item discrimination index (the value of B for the correct alternative). It is very unlikely that a good item would have a value for B that is noticeably negative, because that would mean that a greater proportion of the low-scoring group got the item correct than the high-scoring group. Therefore, if B is noticeably negative (say, less than -.20) examine the item carefully, checking especially to see that the item was scored correctly, that it is unambiguous, and that the indicated correct answer is indeed correct.

(f) Look at the values of B for the distractors. If any of them are noticeably positive (say, above .20), check the item to see if it is ambiguous, or if the distractor could possibly be a correct answer.

(g) If either p or B for "omits" is noticeably positive, examine the item for ambiguities. It is assumed, here, that examinees are not being penalized for guessing and, therefore, there is no extrinsic motivation for an examinee not to pick an alternative.

(h) Consider the number of examinees (especially high-scoring examinees) who did not reach the item. If many examinees did not reach it, (e.g., see Table 2.1) the item may be all right, but it is likely that examinees were not allowed enough time when they were tested. Unless a domain-referenced test is intended to be speeded, examinees should have a reasonable amount of testing time. Otherwise, the examinees' scores will not adequately reflect their ability.

The above suggestions should be regarded as reasonable "rules-of-thumb"--not dogmatic directives. No such rules, and no amount of item analysis data, absolve item developers and investigators from employing common sense and good judgment based on experience and content-matter knowledge.

Other Considerations

In norm-referenced testing contexts it is not uncommon for items to be discarded or revised if the value of a discrimination index is positive but small. This criterion should not be used in domain-referenced testing contexts. Indeed, frequently in such contexts many good items are virtually guaranteed to have positive but small values for a discrimination index. Also, in norm-referenced testing contexts a high discrimination index is frequently viewed almost as an indicator of an ideal item. This perspective should not be taken in domain-referenced testing contexts--at least not in the sense that highly discriminating items are preferred over moderately discriminating ones. In domain-referenced testing situations, emphasis is placed upon content, and discrimination indices should be used solely as an aid in identifying flawed items--not a basis for classifying items into degrees of quality.

In an ideal world, all items in the universe would undergo item analysis before any decisions were made about examinees based on any items in the universe. This ideal is seldom feasible in practice. Even so, no item should be used as a basis for making decisions about examinees until it has been subjected to an item analysis. To address this issue the following procedure can be used. First, in the initial stages of developing a universe of items, prior to using the items for decision-making, a reasonably large sample of them should undergo item analysis using a representative group of examinees. Items that do not successfully clear this hurdle should be discarded or revised. Second, to gather item analysis data on other available items, or items subsequently developed, one can include a small number of them in operational versions of domain-referenced tests. However, examinee scores on any such additional item should not be used as part of the examinee total scores for decision-making--at least not until the item analysis data have been studied to verify that the item has no obvious flaws.

If the above approach is taken of including new items with old items in a domain-referenced test, then it is important that the investigator not confuse the total number of "scored items" (those not undergoing item analysis) and the total number of items physically in the test. Elsewhere in this handbook, when test length, $\underline{n}$, is discussed it is always assumed that $\underline{n}$ is the total number of items excluding those (if any) undergoing item analysis.

As discussed above, conducting an item analysis usually involves classifying examinees into groups based on total test score. If new items are included with old items, then total test score should be based on the old items only. Of course, in the initial stages of constructing a universe, or pool of items, total test score will have to be based on new items only. In either case, the investigator must choose a range of scores associated with each group. Seldom can this decision be made in a completely unambiguous manner, because a firm basis for this decision would necessitate information that is seldom available at the time the decision needs to be made. For example, in initial stages of universe construction, a cutting score may not have been firmly established. Furthermore, as will be discussed later, even under the best of circumstances, it is impossible to assign examinees to groups in a manner that is guaranteed to be completely devoid of error. Even so, for item analysis purposes a firm basis for assigning examinees to groups is not absolutely necessary--good informed judgment based on experience is generally sufficient.

The above discussion of item analysis procedures has been couched in terms of multiple-choice items. For free-response items the procedure and guidelines are essentially the same. The principal differences are that: (a) a free-response item can be viewed as an item with two alternatives--correct and incorrect; and (b) the investigator needs to study all examinee responses to make sure that all correct responses have been identified.

3. Establishing a Cutting Score

One of the initial tasks typically encountered by an investigator in a domain-referenced testing environment is to establish a cutting score, π_o , expressed as a proportion of items correct for the universe of items. Of course, π_o is not required if mastery type decisions are not going to be made and interest is restricted to estimating an examinee's universe score. However, in most domain-referenced testing situations, mastery type decisions are made and, consequently, a cutting score is required.

On rare occasions there is a known relationship between examinee performance on the universe of items (or a large part of the universe) and some external criterion such as on-the-job performance or performance in some subsequent level of instruction. Such data are indeed rare, however, because they are usually very difficult to obtain. For example, if some measure of on-the-job performance is viewed as a criterion, then one would have to take the following steps to obtain the data required to use such performance as a basis for establishing a cutting score: (a) test a representative group of examinees using a large number of items from the universe; (b) allow all these examinees, including those with low scores, to undertake the job under consideration; and (c) evaluate the performance of each of these examinees on the job. Three problems are usually encountered in attempting to carry out these steps. First, these steps are usually time-consuming and expensive.

Second, it is frequently considered undesirable (and sometimes ethically unacceptable) to allow low-scoring examinees to undertake the job in question. And third, usually the evaluation of on-the-job performance is both difficult and subject to considerable error.

For these reasons, among others, external criteria are seldom used (at least directly) in the process of establishing a cutting score for domain-referenced testing purposes. Rather, it is common for a cutting score to be defined based upon the judgments of raters, judges, or experts who are content matter specialists. Of course, such judgments are likely (indeed hopefully) to be influenced by raters' knowledge about potential external criteria and about how persons generally perform on such criteria. However, such information is not usually quantified directly. Rather several procedures exist for eliciting from raters their beliefs about how minimally competent persons would perform on the universe of items, the argument being that such judgments provide a basis for establishing a cutting score π_o that separates mastery (or probably acceptable performance) from non-mastery (or probably unacceptable performance).

Procedure

In one procedure for establishing a cutting score, each of a set of raters, judges, or content matter specialists is asked to provide an independent assessment of the probability that a minimally competent examinee would get each item correct. The average probability over raters and items (called $\bar{y}$ below) is frequently used as the cutting score π_o , and various statistics can be calculated to assess how variable this average probability would be if the study were replciated a large number of times. Knowledge about such variability is important in revealing the extent to which raters agree in their judgments about what cutting scores should actually be established.

Using this procedure data are collected in the following manner:

(a) A group of _t_ raters, and a sample of _m_ items from the universe, are identified where _t_ and _m_ are as large as time and other constraints will allow;

(b) Each rater is told to provide, for each item, a probability reflecting that rater's belief about the likelihood that a minimally competent examinee would get that item correct;

(c) Items are presented to each rater in a random order--the important point being that the items are ordered differently for each rater;

(d) Each rater works independently of every other rater (i.e., raters do _not_ discuss their judgments with each other); and

(d) Raters are told to report their probabilities in units of 1/10 (i.e., the probabilities that might be assigned are 0.0, 0.1, 0.2, . . ., 1.0).

Table 3.1 reports a set of data that might result from such a study with t = 5 raters and m = 20 items. These numbers are relatively small solely for the purpose of simplifying subsequent illustration of computations. An entry in the body of Table 3.1 is denoted y_{ri} , the probability assigned by a rater r to an item i. (The symbol y is used here to distinguish these probabilities from examinee scores on a test, which are later denoted with the symbol x.) Along with the probabilities, Table 3.1 reports means, variances, and standard deviations. For example,

(a) an entry in the row labeled $\bar{y}_r$ is the mean probability assigned to items by rater r, and $\hat{s}(\bar{y}_r) = .083$ is the standard deviation (across raters) of these rater mean probabilities;

(b) an entry in the column headed $\bar{y}_i$ is the mean probability assigned to item i, and $\hat{s}(\bar{y}_i) = .086$ is the standard deviation (across items) of these item mean probabilities;

(c) an entry in the row labeled $\hat{s}(y_{ri})$ is the standard deviation of the probabilities assigned to items by rater r; and

(d) $\bar{y} = .80$ is the mean probability over all 20 items and all 5 raters.

In a cutting score study, interest is usually focused principally on $\bar{y}_r$ and $\bar{y}$. We may call $\bar{y}_r$ the "cutting score assigned by rater r" because it reflects that rater's belief about the proportion of items that a minimally competent examinee would get correct. Similarly, we may call $\bar{y}$ the "study cutting score," and as such it is, in a certain statistical sense, the best value to choose for π_o.

Table 3.1

Establishing a Cutting Score:

A Synthetic Data Set with Sample Statistics

Item	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	$\bar{y}_i$
1	0.8	0.9	0.8	0.7	0.6	0.76
2	0.8	0.7	0.9	0.7	0.5	0.72
3	0.9	1.0	0.9	0.8	0.6	0.84
4	0.6	0.8	0.8	0.9	1.0	0.82
5	1.0	0.9	0.9	0.9	0.7	0.88
6	0.9	1.0	1.0	0.8	0.7	0.88
7	0.6	0.8	0.5	0.6	0.6	0.62
8	0.8	0.7	0.7	0.6	0.5	0.66
9	0.7	0.8	0.8	0.7	0.5	0.70
10	0.8	0.9	0.7	0.8	0.6	0.76
11	0.7	1.0	1.0	0.9	1.0	0.92
12	0.9	1.0	0.8	0.9	0.6	0.84
13	0.7	0.8	0.9	0.9	1.0	0.86
14	0.9	1.0	0.8	0.9	0.7	0.86
15	1.0	0.9	0.9	0.9	0.6	0.86
16	0.9	0.9	0.8	0.8	0.6	0.80
17	0.9	0.8	0.7	0.8	0.6	0.76
18	0.9	1.0	0.9	0.8	0.6	0.84
19	0.9	1.0	1.0	1.0	0.7	0.92
20	0.7	0.7	0.8	0.8	0.5	0.70
$\bar{y}_r$	.82	.88	.83	.81	.66	$\bar{y} = .80$
$\hat{s}^2(y_{ri})$	.0143	.0122	.0148	.0115	.0257	$\hat{s}^2(\bar{y}_i) = .0074$
$\hat{s}\ (y_{ri})$	.120	.111	.122	.117	.160	$\hat{s}\ (\bar{y}_i) = .086$
	$\hat{s}^2(\bar{y}_r) = .0069$			$\hat{s}(\bar{y}_r) = .083$		

It is evident from the values of $\bar{y}_r$ in Table 3.1 that Raters 1, 3, and 4 are in reasonably close agreement concerning choice of a cutting score, but Rater 2 thinks the cutting score should be higher than 0.80 and Rater 5 thinks it should be considerably lower than 0.80. This disagreement among raters is reflected in the quantity $\hat{s}(\bar{y}_r) = .083$. Such disagreement is not unusual and probably should be expected because even well-qualified raters may have different opinions about minimal competence and/or the relationships between minimal competence and the items used in the study. Indeed, one purpose of a cutting score study is to reveal such differences of opinion in a systematic and objective manner.

Variability in Study Cutting Scores

For the purpose of examining variability in $\bar{y}$, $\hat{s}(\bar{y}_r)$ is relevant but not actually the quantity of principal interest. Rather, one would ideally like to know how variable $\bar{y}$ would be if the study were replicated (under similar conditions) a large number of times. Let us describe this variability in $\bar{y}$ in terms of a standard deviation and identify it as $\sigma(\bar{y})$. Clearly, if $\sigma(\bar{y})$ were small, then, even if raters disagreed to some extent concerning the cutting score resulting from a single study, such disagreement would not seriously impact one's confidence in using $\bar{y}$ as a cutting score. However, if $\sigma(\bar{y})$ were large, then one might want to keep this fact in mind when making decisions based on $\bar{y}$.

Even though there is typically only one cutting score study available, it is still possible to estimate the standard deviation of $\bar{y}$ that would result if the study were replicated a large number of times. Table 3.2 reports three such estimates along with their numerical values for the data in Table 3.1. These three estimates are similar in that each of them assumes that each (hypothetical) replicated study involves a different sample of _t_ raters (t=5 in Table 3.1) and a different sample of items. As described below, the three estimates differ with respect to the number of items involved in each replicated study: Equation 3.1 in Table 3.2 would be appropriate if an investigator wanted to consider (hypothetical) replicated studies involving _m_ items--the same number of items used in the actual cutting score study. Under this circumstance, Table 3.2 shows that $\sigma(\bar{y}) = .041$ for the data in Table 3.1. If, however, an investigator wanted $\sigma(\bar{y})$ over replicated studies involving _n_ items--a number different from (usually smaller than) _m_, then the appropriate estimate would be obtained from Equation 3.2 in Table 3. For example, given the synthetic data and a test length of n=10 items, Table 3.2 shows that $\sigma(\bar{y}) = .045$.

A third estimate of $\sigma(\bar{y})$ is obtained by assuming that replicated studies would each involve rating _all_ items in the universe. Under this circumstance, the appropriate estimate of $\sigma(\bar{y})$ is Equation 3.3 in Table 3.2; and for the synthetic data $\sigma(\bar{y}) \approx 0.036$. This value is less than either of the other two estimates of $\sigma(\bar{y})$ because $\sigma(\bar{y})$ _decreases_ as the number of items _increases_.

Table 3.2

Equations and Illustrative Computations for Determining the Standard Deviation of a Mean Cutting Score

Equation	Computations Using Data in Table 3.1
Let t = number of raters used in study	$t = 5$
m = number of items used in study	$m = 20$
Define $A = \dfrac{1}{m(t-1)}\left[\begin{pmatrix}\text{Average value} \\ \text{of } \hat{s}^2(y_{ri})\end{pmatrix} - \hat{s}^2(\bar{y}_i)\right]$	$A = \dfrac{1}{(20)(4)}\left[\dfrac{1}{5}\begin{pmatrix}.0143 + .0122 + .0148 \\ + .0115 + .0257\end{pmatrix} - .0074\right] = .0001$
Standard deviation of $\bar{y}$ over different studies using t raters and m items:	
(3.1) $\sigma(\bar{y}) = \sqrt{\hat{s}^2(\bar{y}_r)/t + \hat{s}^2(\bar{y}_i)/m - A}$	$\sigma(\bar{y}) = \sqrt{(.0069)/5 + (.0074)/20 - .0001} = .041$
Standard deviation of $\bar{y}$ over different studies using t raters and some number of items, n, different from m:	If n = 10 items
(3.2) $\sigma(\bar{y}) = \sqrt{\hat{s}^2(\bar{y}_r)/t + \hat{s}^2(\bar{y}_i)/n - A}$	$\sigma(\bar{y}) = \sqrt{(.0069)/5 + (.0074)/10 - .0001} = .045$
Standard deviation of $\bar{y}$ if each of the t raters rated all items in the universe:	
(3.3) $\sigma(\bar{y}) = \sqrt{\hat{s}^2(\bar{y}_r)/t - A}$	$\sigma(\bar{y}) = \sqrt{(.0069)/5 - .0001} = .036$

Any one of these estimates might be of interest to an investigator; however, the third estimate is especially relevant for many (if not most) domain-referenced testing situations. Recall that a cutting score is defined as a proportion of items correct for the universe of items. It follows that ideally one would like to have each rater rate every item in the universe to obtain each of the "rater cutting scores." It is almost always impossible to obtain such data directly, but even so Equation 3.3 allows us to estimate $\sigma(\bar{y})$ under this circumstance. This equation is also appropriate if the rating procedure is followed for all items that occur in each and every form of a domain-referenced test.

One particular use of $\sigma(\bar{y})$ in Equation 3.3 is in establishing a confidence interval for the cutting score. For example if one goes one standard deviation to the right and left of $\bar{y}$, then one obtains a 68% confidence interval for the cutting score π_o. For the synthetic data this interval extends from

$$\bar{y} - \sigma(\bar{y}) \quad = \quad .800 - .036 \doteq .76$$

to

$$\bar{y} + \sigma(\bar{y}) \quad = \quad .800 + .036 \doteq .84,$$

and this interval is represented (.76, .84). In words, we can say that if the cutting score study were replicated a large number of times (each time using all items in the universe), about 68% of the time we would expect to obtain values of $\bar{y}$ between .76 and .84.

Given these data, therefore, in a certain statistical sense $\bar{y}$ = .80 is the best single number (proportion of items correct) to use as a cutting score, π_o; however, an investigator is well advised to entertain some uncertainty about whether or not this value for π_o is "correct" in some absolute sense. Also, as will be indicated in Section 4 for some purposes, procedures are available that employ what is called an "indifference zone" for the cutting score π_o; and the confidence interval discussed above can be helpful in picking an indifference zone.

Other Considerations

One factor that can contribute greatly to differences among raters in their $\bar{y}_r$ values is differential ideas about what constitutes minimal performance. Any definition of minimal competence is almost always a matter of judgment (packing a parachute may be an exception!), but very disparate notions about minimal competance can render a cutting score study of relatively little value. At the same time, however, the raters themselves should be well qualified to define what minimal competence is, or at least to have a voice in any such definition. In particular, it is very difficult, if not impossible, for raters to participate in a cutting score study using someone else's definition of minimum competence. For these reasons, it is advised that raters have the opportunity to discuss their possibly different notions about minimal competence prior to conducting the actual study. Hopefully, they can reach some consensus or at least mitigate their differences of opinion in a mutually acceptable manner.

Another issue to be considered is the manner in which items are provided to raters--specifically, are the answers provided along with the items? All things considered, it is probably best that answers be supplied. In doing so, one can obtain an additional check on thc correctness of the indicated answers, and raters are probably more likely to pay careful attention to each item individually. Assuming that the answers are supplied, each rater should be directed to indicate any items that he/she judges to be keyed incorrectly. If it is determined after the raters complete their task that an item is keyed incorrectly, it (and the probabilities assigned to it) should be eliminated from the study, and the item should be revised or discarded. If, on the other hand, it is determined after careful consideration that a rater said an item was keyed incorrectly, but actually it was keyed correctly, then that rater's judgment (i.e., assigned probability) for that item should be eliminated in determining $\bar{y}$. This can happen-- each individual rater is not infallible, even in his/her area of expertise.

Table 3.1 illustrates the rather common occurrence of one rater (in this case Rater 5) providing judgments that are markedly different from the judgments provided by other raters. Even so (assuming all raters were chosen carefully in the first place), an atypical rater should not be eliminated from the study unless there is an obvious reason (e.g., sickness) for that rater's atypical judgments. If such a reason exists, then all statistics should be re-calculated based on the reduced set of raters. [For example, if Rater 5 were eliminated from the synthetic data, then the reader can verify that $\bar{y} = .835$; $\hat{s}(\bar{y}_r) = .031$; and, using Equation 3.3, $\hat{\sigma}(\bar{y}) = .021$.]

One modification of (or addition to) this procedure for establishing a cutting score involves having the raters, as a group, provide a consensus probability for each item after they have independently provided their judgments about each item. Then the mean of these consensus probabilities is used as the cutting score. If this modification is employed, the resulting data should be examined very carefully to ensure that no single rater is exerting undue influence over the judgments of other raters. (Also, if this modification is used one should keep in mind that forced consensus is not really agreement, although forced consensus can effectively hide disagreement.)

4. Establishing an Advancement Score

When domain-referenced testing is employed to make mastery/non-mastery types of decisions, it is necessary to consider a cutting score, π_o; but, in addition, the investigator must specify an observable score, x_o, such that an examinee who gets x_o or more items correct will be declared a master; and an examinee who gets fewer than x_o items correct will be declared a non-master. This score is called an advancement score, with the symbol x_o referring to the advancement score in terms of number of items correct and (later) the symbol c_o referring to the advancement score in terms of proportion of items correct.

In principle, one wants to pass, or advance, an examinee if that examinee's universe score, π_p, is equal to or greater than the cutting score, π_o. However, one cannot directly use such a decision rule because a specific domain-referenced test will consist of only a sample of items from the universe. Based on any sample of items, an examinee's observed mean score, $\bar{x}_p$, can be calculated. but not the examinee's universe score, π_p. Furthermore, the cutting score, π_o, may not correspond with a possible observed mean score for test of n items. (For example, if n = 10, then no proportion of items correct will correspond with a cutting score of .85.)

Let us suppose that, as a result of some cutting score study, π_o is specified to be .80, and let us assume that a test will consist of $n = 10$ items. Since .80 x 10 = 8, an investigator might decide that the advancement score should be:

$x_o = 8$ in terms of number of items correct; or

$c_o = x_o/n$

$= 8/10$

$= .80$ in terms of proportion of items correct.

In this example, choosing x_o to be eight items correct may appear reasonable and, indeed, this particular advancement score may be a good choice in some particular context. However, the "logic" presented above for choosing an advancement score is rather superficial. For example, this logic does not take into account the fact that an observed score may be, and usually is, different from a universe score. As will become evident later, a more thorough analysis could lead to choosing some advancement score other than $x_o = 8$.

The purpose of this section is to provide a reasonably sound, yet relatively simple, table-look-up procedure for choosing an advancement score. Even though this procedure is quite simple compared to others that might be used, it does involve consideration of several technical issues. Specifically, to use this procedure, one must first specify a test length, a loss ratio, and an indifference zone. These issues are discussed below, followed by an illustration of how to use the table look-up-procedure.

Related Issues

Sometimes, choosing a test length (n) is a more difficult problem than it may appear to be at first glance. All other things being equal, longer tests are to be preferred over shorter tests, because longer tests reduce certain types of errors (discussed more fully later). Also, longer tests are more valid in the sense that they provide a more thorough representation of the intended universe of items. At the same time, however, in domain-referenced testing environments, factors such as available testing time frequently make it very difficult and/or costly to use tests that are very long. For now, it will be assumed that there already exists some reasonable basis for choosing a particular test length, at least for the initial form(s) of a domain-referenced test. In subsequent sections, as different concepts and procedures are developed, it will be possible to identify some reasonable statistics to consider in choosing, or modifying, test length.

Classification errors and loss ratio. The concept of a loss ratio involves a consideration of errors that can be made in *classifying* an examinee as a passing examinee (master) or a failing examineee (non-master). Specifically, there are two classification errors that can be made:

(a) a *false positive* error occurs if an examinee is declared a master (i.e., advanced) who has a universe score *below* π_o; and

(b) a *false negative* error occurs if an examinee is declared a non-master (i.e., not advanced) who has a universe score *above* π_o.

These two classification errors are considered more fully in Section 5 in the context of decisions about individual examinees. Here, our concern is with a certain kind of judgment about false positive and false negative errors. Specifically, in this handbook the term "loss ratio" refers to a number reflecting judgment about the seriousness of a false positive error compared to the seriousness of a false negative error. For example, if false positive errors were judged to be twice as serious as false negative errors, then the loss ratio would be two; and, if both types of classification errors were equally serious, then the loss ratio would be one.

By definition, the specification of a loss ratio involves subjective judgment on the part of a person (or persons) intimately familiar with the testing context. In making this judgment one needs to consider the consequences of inappropriately passing or inappropriately failing an examinee. For example, in many domain-referenced testing contexts, it is frequently argued that an examinee who is inappropriately advanced (false positive error) is likely to be unsuccessful on-the-job or in subsequent instruction; and, this type of error is judged more serious than the time and cost involved in inappropriately re-cycling an examinee through an instructional sequence (false negative error). These particular judgments suggest that a loss ratio, in such contexts, should be defined as some number greater than one--perhaps two, but probably not three unless instructional time and cost are quite unimportant.

Indifference zone. An indifference zone is some range of universe scores within which one is "indifferent" about false positive and false negative errors. Let us identify the lower limit of this range as π_L , the upper limit as π_H , and the range itself as (π_L , π_H). Suppose an investigator is able to specify values for π_L and π_H such that, for any examinee whose universe score is between π_L and π_H , there is virtually no loss involved in declaring a true master to be a non-master or in declaring a true non-master to be a master. In such a case the interval (π_L , π_H) may be viewed as an indifference zone. This rather direct approach to defining an indifference zone may or may not make sense in a particular context.

Another approach involves the procedure for establishing a cutting score discussed in Section 3. Specifically, consider again $\sigma(\bar{y})$ in Equation 3.3, which is the standard deviation of $\bar{y}$ over replicated studies, if each study involved all the items in the universe. It was stated in Section 3 that $\bar{y}$ can serve as π_o and a 68% confidence interval for π_o can be viewed as extending from $\bar{y} - \sigma(\bar{y})$ to $\bar{y} + \sigma(\bar{y})$, approximately. This confidence interval (or something close to it) might be viewed as an indifference zone. Consider, for example the synthetic data treated in Section 2. For these data, $\bar{y} = .80$; using Equation 3.3, $\sigma(\bar{y}) = .036$; and the 68% confidence interval is (.76 to .84). Since this interval indicates a degree of uncertainty about some "ideal" value for a cutting score, it seems reasonable to assume that an investigator might have little basis for being anything but indifferent about classification errors for examinees whose universe scores lie in the interval (.76 to .84).

In considering either of the above approaches to establishing an indifference zone, it needs to be recognized that these procedures are not to be viewed as statistical excuses for being indifferent, in the sense of uncaring, about individual examinees who have observed mean scores close to π_o. Rather, these procedures are to be viewed as aids in the process of establishing an indifference zone, which is a necessary consideration for picking an advancement score using the table discussed below.

Advancement Score Table

Given a test length, a loss ratio, and an indifference zone, Table A.1 provides a specific advancement score, x_o , in terms of number of items correct. (To obtain the advancement score in terms of proportion of items correct, one simply uses the relationship $c_o = x_o/n$.) The rows of Table A.1 are associated with different test lengths, ranging from 6 to 30 items; and the columns are associated with 20 indifference zones, organized according to the mid-points of the zones, with mid-points ranging from .65 to .90. For each row and column, there are three tabled entries (separated by slashes) corresponding to advancement scores associated with loss ratios of 1, 2, and 3, respectively.

To illustrate use of Table A.1, let us consider the following judgments about test length, loss ratio, and indifference zone:

(a) Test length. Let us assume that testing time is at a premium, and the universe of items is rather narrow. Taking these two considerations into account, it is judged that about n = 10 test items seems reasonable.

(b) Loss ratio. Let us assume that the domain-referenced testing context is one in which false positive errors are judged to be somewhat more serious than false negative errors, and a loss ratio of about two seems reasonable.

(c) Indifference zone. Let us suppose that it is decided to use the results of a cutting score study in making judgemnts about an indifference zone. Specifically, let us suppose that the results reported in Section 2 are based on the appropriate universe of items. This study suggests that an approximate 68% confidence interval for π_o is (.76 to .84); and it will be assumed that this confidence interval can serve as an approximate indifference zone.

Now, given the above judgements, to pick an advancement score, one uses the fifth row (n = 10) and second column(.75 to .85) of the second page of Table A.1. The tabled entries corresponding to this row and column are 9/9/9. Since all of these entries are the same number, it is obvious that the advancement score is $x_o = 9$ or $c_o = 9/10 = .90$. To be specific, since the loss ratio has been defined as two, the second entry is actually the advancement score for this illustration.

In the above example, note that the indifference zone (.75 to .85) specified in the second column of the second page of Table A.1 is not exactly equal to the indifference zone of (.76 to .84), which was initially chosen. Any such slight disparity can be overlooked without serious consequences, because, for the most part, the procedure used to develop Table A.1 is insensitive to small disparities in indifference zones. Furthermore, it is not necessary that π_o be exactly at the midpoint of the indifference zone. Indeed, for reasons beyond the scope of this handbook, it is sufficient that π_o be somewhere within the indifference zone.

Table A.1 indicates (and the above example illustrates) that this procedure for choosing an advancement score is also relatively insensitive to small changes in loss ratio. Indeed, for any specific test length and indifference zone in Table A.1, the suggested advancement scores differ by at most, one correct item.

The above points about "insensitivity" have been made to highlight the fact that this procedure for choosing an advancement score does not necessitate arguing about minute differences of opinion with respect to an appropriate indifference zone or loss ratio--a reasoned consideration of these issues is sufficient for the procedure.

5. Errors of Measurement, Errors of Classification, and Inferences about an Examinee's Universe Score

Sections 2, 3, and 4 have considered issues that are addressed prior to making any decision about an examinee. Let us now assume that the issues discussed in Sections 2, 3, and 4 have been addressed, a domain-referenced test of n items has been administered to a group of examinees, and each examinee's score on the test has been determined. In this section, consideration is given to the precision, or quality, of certain statements, or decisions, that might be made about an examinee. To address these issues, the only examinee datum that will be employed is the examinee's test score. To simplify notation in this section, usually the examinee's number of items correct will be denoted x, the examinee's proportion of items correct will be denoted $\bar{x}$ (rather than $\bar{x}_p$), and the examinee's universe score will be denoted π (rather than π_p).

It cannot be emphasized enough that π is always unknown, and $\bar{x}$ is only an estimate of π. Consequently, there is always some degree of uncertainty about any statement concerning π. For example, if $\bar{x} = .80$, one may say that π is "about" .80, but this statement clearly suggests that π and $\bar{x}$ may be different, and perhaps dramatically different. This difference between $\bar{x}$ and π is called an error of measurement.

Furthermore, since $\bar{x}$ is an imperfect estimate of π, mastery/non-mastery decisions based on $\bar{x}$ (or x) may be incorrect, and an error of classification may be made. This issue was introduced in the previous section in the context of specifying a loss ratio. In this section, errors of classification are considered in more detail, from the perspective of decisions about examinees.

It needs to be recognized that, since π is unknown, one cannot specify whether or not a classification error has been made for an individual examinee; nor, can one specify a particular value for an individual examinee's error of measurement. However, given n and $\bar{x}$ (or x), it is possible to make statements about the probability of correct and incorrect decisions, and about likely values of π. Procedures for doing so are described and illustrated in this section, after a more detailed consideration of errors of measurement and classification.

Errors of Measurement and Classification.

Recall that an examinee's universe score is the porportion of items, π, that the examinee would get correct if the examinee were administered all items in the universe. Suppose an examinee takes a domain-referenced test with $n = 10$ items and gets $x = 8$ items correct. It should be intuitively obvious that this does not necessarily mean that the examinee's universe score is $\bar{x} = x/n = 8/10 = .80$. After all, the examinee was tested with 10 items, only; and it is to be expected that $\bar{x} = .80$ is an imperfect estimate of the examinee's universe score. This imperfection in measurement is called measurement error. Specifically, measurement error is the difference between an exmaminee's test score (expressed as a proportion of items correct, $\bar{x}$) and the examinee's universe score:

$$\Delta = \bar{x} - \pi.$$

Note the use of the symbol Δ to designate measurement error. Clearly, Δ can be either positive or negative, as well as being either large or small.

It is evident from the definition of Δ that a cutting score, π_o, plays no role in considerations regarding error of measurement. However, for mastery/non-mastery decisions a cutting score, π_o, is involved; and for such decisions, an error of classification may be made in addition to an error of measurement. As noted in Section 4, there are two types of errors of classification:

(a) a false positive error (f+) occurs if an examinee is declared a master ($x \geq x_o$) when the examinee's universe score is below π_o; and

(b) a false negative error (f-) occurs if an examinee is declared a non-master ($x < x_o$) when the examinee's universe score is at or above π_o. These two possible errors of classification are represented in Table 5.1 along with the two possible correct decisions--namely, passing an examinee who has a universe score at or above π_o (c+), and failing an examinee who has a universe score below π_o (c-).

To better appreciate errors of measurement and classification, consider Figure 5.1 in which it is assumed that $\pi_o = .80$, $n = 10$, and $c_o = .90$. For 12 pairs of values for $\bar{x}$ and π, Figure 5.1 represents the resulting error of measurement and error of classification or correct decision. As illustrated in Figure 5.1:

(a) a false positive decision implies that a positive error of measurement ($\bar{x} > \pi$) is involved (see lines G, H, and I in Figure 5.1);

(b) a false negative decision implies that a negative error of measurement ($\bar{x} < \pi$) is involved (see lines J, K, and L in Figure 5.1); and

(c) even when a correct (positive or negative) decision is made, an error of measurement (positive or negative) may be involved (see lines A-F in Figure 5.1).

In short, the occurrence of an error of measurement does not necessarily mean that an error of classification will be made; however, an error of classification is always associated with an error of measurement, and

Table 5.1

Correct Mastery/Non-Mastery Decisions and

Errors of Classification

Observed Score	Universe Score	
	$\pi < \pi_o$	$\pi \geq \pi_o$
$x < x_o$ (Fail)	Correct Negative Decision (c-)	False Negative Error (f-)
$x \geq x_o$ (Pass)	False Pcsitive Error (f+)	Correct Positive Decision (c+)

Note. The symbol $>$ means "greater than," the symbol $\geq$ means "greater than or equal to," the symbol $<$ means "less than," and the symbol $\leq$ means "less than or equal to."

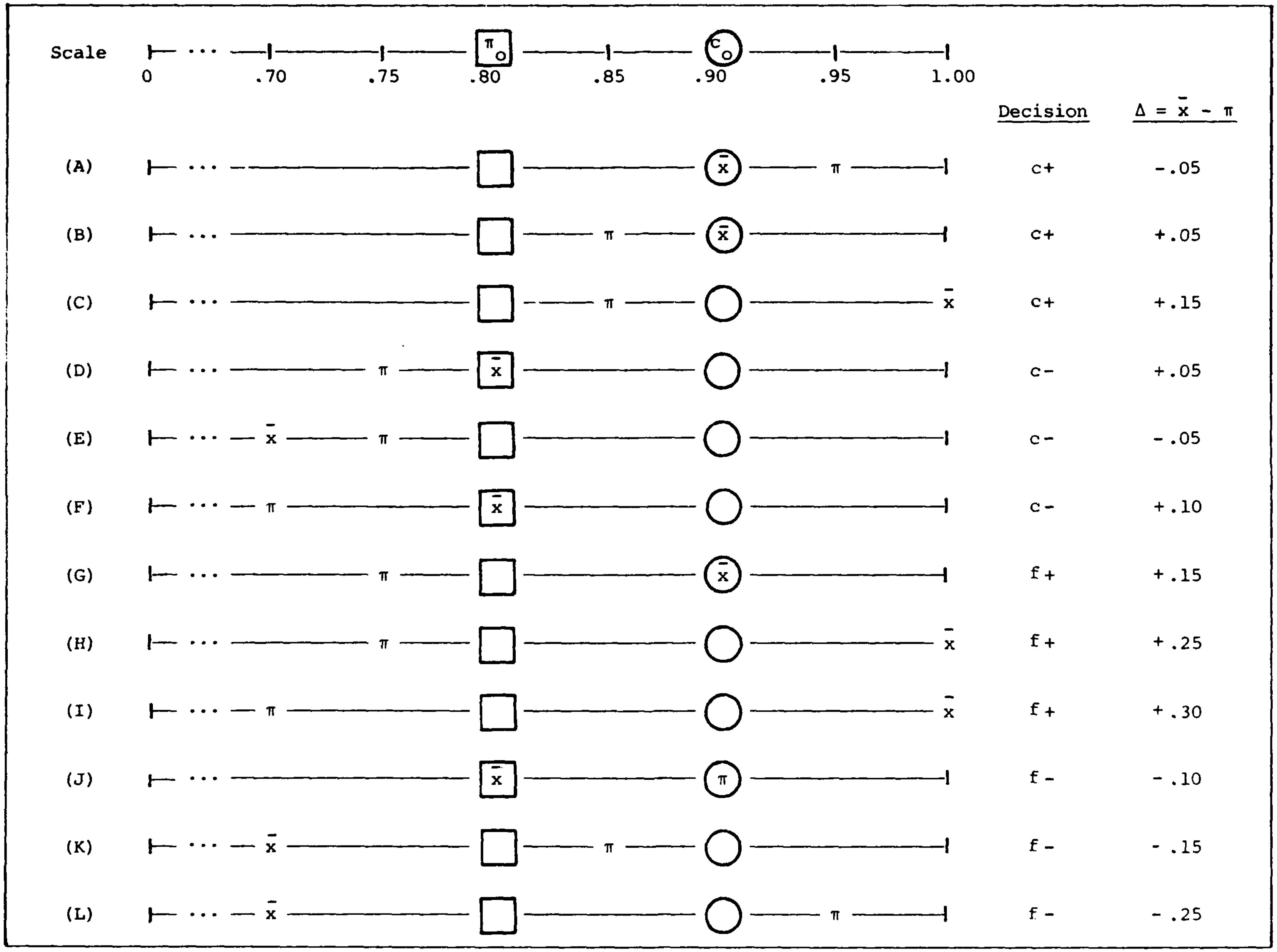

Figure 5.1. Illustration of Errors of Classification and Errors of Measurement.

frequently a rather large error of measurement. Indeed, errors of classification arise because errors of measurement are involved. This is one reason why it is highly advisable to pay attention to issues surrounding errors of measurement--even if the principal focus of domain-referenced testing is mastery/non-mastery decisions.

It should be noted also that, if an error of classification is made, it is not correct to describe the error of classification as being either large or small--such an error is either made or it is not made, nothing more. For example lines G and I in Figure 5.1 both represent false positive classifications errors, and line G does not represent a larger classification error than line I. Rather, line G represents a larger error of measurement than line I.

It needs to be recognized that, since an individual examinee's universe score is unknown, we cannot directly determine the error of measurement for an individual examinee. For the same reason, it is impossible to say, for certain, whether or not a classification error has been made for an individual examinee. However, given n and x (or $\bar{x}$) it is possible to make statements about: (a) probabilities associated with correct and incorrect decisions; and (b) likely values for π. Procedures for doing so are treated in the next two parts of this section.

Probabilities of Correct and Incorrect Decisions

Since one cannot say, for certain, whether or not a classification error has been made for an individual examinee, it is reasonable to ask, "How probable is it that an examinee with a score of x (or $\bar{x}$) on an n-item test has been misclassfied?" Technically, there are many answers to this question, depending on the assumptions one is willing to make. The approach taken here to answering this question involves using Table A.2 which was developed under very simple assumptions (see Appendix B). Roughly, speaking, these assumptions imply that all we know about an examinee is the examinee's test score, and the fact that the examinee took a test consisting of a sample of n items from a large universe of items.

Table 5.2 provides a step-by-step procedure, with examples, for determining probabilities associated with correct and incorrect decisions. This procedure involves nothing more complicated than identifying an entry in Table A.2 and possibly subtracting it from 100. Note that, in this handbook, a probability is usually identified and discussed as a percent ranging from *0 to 100. This convention has been adopted to avoid confusing* a statement about a probability with a statement about an examinee's universe score (π) or observed mean score ($\bar{x}$), both of which range from 0 to 1.

It is suggested that, whenever mastery/non-mastery decisions are to be made, the investigator examine the probabilities in Table 5.2--at least the probabilities of incorrect decisions for examinees near the cutting score. For example, using the procedure in Table 5.2 with n = 10, $\pi_o = .80$, and $c_o = .90$,

Prob (f-) = 5% if x = 6,

Prob (f-) = 16% if x = 7,

Prob (f-) = 38% if x = 8,

Prob (f+) = 32% if x = 9, and

Prob (f+) = 9% if x = 10.

Table 5.2

Use of Table A.2 to Determine Probabilities of Correct and Incorrect Classification Decisions

Procedure and Equations	Examples
Using the left-hand side of Table A.2: (a) locate the row for n and x, (b) locate the column for π_o , and (c) let TE be the tabled entry at the intersection of this row and column. Then, (d) if $x < x_o$ use Equations 5.1 and 5.2, or (e) if $x \geq x_o$ use Equations 5.3 and 5.4.	Suppose n = 10, $\pi_o = .80$, and $x_o = 9$ Example 1: If x = 7 (i.e., $\bar{x} = .70$, then the tabled entry is TE = 16; and, since $x < x_o$, Equations 5.1 and 5.2 are used. Example 2: If x = 9 (i.e., $\bar{x} = 90$), then the tabled entry is TE = 68; and since $x \geq x_o$, Equations 5.3 and 5.4 are used.
Probability of a Correct Negative Decision:	
(5.1) Prob (c-) = (100 - TE)%	Example 1: Prob (c-) = (100 - 16)% = 84%
Probability of a False Negative Decision:	
(5.2) Prob (f-) = TE%	Example 1: Prob (f-) = 16%
Probability of a Correct Positive Decision	
(5.3) Prob (c+) = TE%	Example 2: Prob (c+) = 68%
Probability of a False Positive Decision:	
(5.4) Prob (f+) = (100 - TE)%	Example 2: Prob (f+) = (100 - 68)% = 32%

Given such results, an investigator might decide that, when $x = 8$ or 9 the probability of an incorrect decision is unacceptably large. If so, the investigator might consider retesting examinees with scores of 8 or 9 using a different sample of items.

Suppose, for example that an examinee got 8 out of 10 items correct, initially, and 10 out of 10 items correct on a retest. The cutting score is still $\pi_o = .80$; but, over both tests,

$$x = 8 + 10 = 18 \qquad \text{and} \qquad n = 10 + 10 = 20.$$

To make a decision about this examinee, the investigator must recognize that the effective test length for this examinee is now $n = 20$; and, consequently, a new value for the advancement score, x_o , must be determined using the procedure discussed in Section 4. Suppose that x_o turns out to be 17 (which is the value of x_o when the loss ratio is two and the indifference zone is .75 to .85). Since $x = 18$ is greater than $x_o = 17$, the examinee should be advanced; and Table A.2 indicates that, under these circumstances, the probability of a false positive error is 18%.

The probabilities of correct and incorrect decisions resulting from the procedure outlined in Table 5.2 do not depend on having examinee scores on a specific test; rather, these probabilities are for any test consisting of a sample of 10 items from a very large universe. It follows that an investigator might consider making a decision about test length based on an examination of probabilities of incorrect decisions, for tests of different length. In Section 6 a closely related issue is treated in detail.

Intervals for an Examinee's Universe Score

Even though it is impossible to specify a numerical value for error of measurement for an individual examinee, it is possible to make statements about probable values of π, given n and x (or $\bar{x}$). More specifically, it is possible to determine:

(a) the probability that π is between two particular values (π_1 and π_2) specified by the investigator; and

(b) an interval (or range of values) for π such that the investigator can say with P% certainty that the examinee's universe score is within the interval.

A procedure for determining the probability referenced in (a), above, and the interval referenced in (b), above, are provided in Table 5.3. To be techically correct, we should not speak about the probability or the interval because there are many such probabilities and intervals, depending on the assumptions one is willing to make. Since the procedures outlined in Table 5.3 involve a simple application of Table A.2, the assumptions for this procedure are those involved in generating Table A.2 (see previous discussion of Table A.2 and Appendix B).

It should be noted that (a) and (b), above, answer different questions. Specifically, (a) answers the question:

"Given n, x, and two investigator-specified values (π_1 and π_2), what is the probability that π is between π_1 and π_2?"

For example, using the procedure in Table 5.3, when n = 10, x = 8, $\pi_1 = .75$, and $\pi_2 = .85$, there is a 32% probability that π is between .75 and .85.

Table 5.3

Use of Table A.2 to Make Statements about Likely Values for π

Procedure and Equations	Examples
Probability that π is Between π_1 and π_2	Suppose $n = 10$, $\pi_1 = .75$, and $\pi_2 = .85$
Given n, x, π_1 and π_2:	
(a) using the <u>left</u>-hand side of Table A.2, locate the row for n and x;	Example 1: If $x = 7$, then $TE_1 = 29$, $TE_2 = 7$, and Prob $(.75 < \pi < .85) = (29 - 7)\% = 22\%$
(b) let TE_1 be the tabled entry in this row under the column headed π_1; and	
(c) let TE_2 be the tabled entry in this row under the column headed π_2.	Example 2: If $x = 9$, then $TE_1 = 80$, $TE_2 = 51$, and Prob $(.75 < \pi < .85) = (80 - 51)\% = 29\%$
(5.5) Prob $(\pi_1 < \pi < \pi_2) = (TE_1 - TE_2)\%$	
P% Credibility Interval for π	Suppose $n = 10$ and $P\% = 80\%$
Given n, x, and P:	
(a) locate the row for n and x in the right-hand side of Table A.2; and	Example 1: If $x = 7$ A P% Credibility Interval for $\pi = (.51, .85)$
(b) let (π_1 , π_2) be the tabled entry in this row under the column headed P-Percent.	Example 2: If $x = 9$ A P% Credibility Interval for $\pi = (.74, .98)$
(5.6) A P% Credibility Interval for $\pi = (\pi_1 , \pi_2)$ (i.e., there is a P% probability that π is between π_1 and π_2)	

By contrast, (b) answers the question:

"Given n, x, and some desired degree of certainty, (P%), what is *a range of values which probably includes* π?"

For example, given n = 10 and x = 8, Table A.2 reports that:

(1) with 67% certainty π is between .67 and .90;

(2) with 80% certainty π is between .62 and .92; and

(3) with 90% certainty π is between .56 and .94.

Note that if one wants to have a greater degree of certainty about the range within which an examinee's universe score probably lies, then one must tolerate a wider interval. For example, the interval (.56, .94) for 90% certainty is quite a bit wider than the interval (.67, .90) for 67% certainty.

Also, given $\bar{x}$ and some desired degree of certainty, the width of an interval decreases as $\underline{n}$ increases. For example, given *n* = 20 and x = 16, $\bar{x}$ = .80 and from Table A.2 a 67% interval is (.71, .87). This interval is shorter than the corresponding interval (.67, .90) *for n* = 10 *and x* = 8. In this sense one can say that long tests are better than short tests, or, more specifically, longer tests are generally associated with a smaller average error of measurement for examinees. This issue of test length and its relationship with errors of measurement is treated in detail in Section 6.

The intervals reported in Table A.2 are sometimes described as <u>credibility</u> intervals. Specifically, Table A.2 reports 67, 80, and 90 percent credibility intervals associated with observed mean scores of $\bar{x} \geq .50$, for test lengths ranging from 5 to 30 items. Similar results can be obtained for other intervals, other test lengths, and/or other observed mean scores using the procedure outlined in Table 5.4. Actually, an interval obtained using the procedure in Table 5.4 is called a confidence interval rather than credibility interval, and the interpretation of a confidence interval is slightly different from the interpretation of a credibility interval. However, for most practical purposes they can be interpreted in about the same way.

As indicated by the example in Table 5.4, one can say with about 6E percent confidence that an examinee with an observed mean score of .75 on a 20-item test probably has a universe score between .65 and .85. By comparison, consider the "corresponding" 67% credibility interval provided in Table A.2. This credibility interval extends from .65 to .83. Clearly, the two intervals are quite close, but not exactly the same. In general, it is recommended that the credibility intervals in Table A.2 be used whenever possible, and that the procedure in Table 5.4 be used when Table A.2 does not apply. For example, Table A.2 does not provide 95 percent intervals, but the procedure in Table 5.4 can be used to obtain such intervals. (Note, however, that the procedure in Table 5.4 does not apply if $\bar{x}_p = 0$ or 1; and this procedure involves a normality assumption that becomes less tenable as $\bar{x}_p$ approaches either 0 or 1.)

In this author's opinion, in domain-referenced testing, it is usually advisable to determine credibility or confidence intervals for examinee

Table 5.4

Equations and Illustrative Computations for Obtaining Confidence Intervals for an Examinee's Universe Score

Equations and Procedure	Example
Let n = number of items in test	Suppose $n = 20$
$\bar{x}_p$ = examinee's observed mean score	$\bar{x}_p = .75$ (i.e., $x = 15$ items correct)
Step 1: Calculate	
(5.7) $\sigma(\Delta_p) = \sqrt{\dfrac{\bar{x}_p(1-\bar{x}_p)}{n - 1}}$	$\sigma(\Delta_p) = \sqrt{\dfrac{.75(1-.75)}{20 - 1}} = \sqrt{.0099} = .10$
Step 2: A P percent confidence interval for the examinee's universe score extends from	
(5.8) $\bar{x}_p - z\,\sigma(\Delta_p)$ to $\bar{x}_p + z\,\sigma(\Delta_p)$	68 percent confidence interval extends from
where $z = 1.00$ if $P = 68$ (percent)	$.75 - 1.00(.10)$ to $.75 + 1.00(.10) = .65$ to $.85$
$z = 1.15$ if $P = 75$ (percent)	
$z = 1.29$ if $P = 80$ (percent)	95 percent confidence interval extends from
$z = 1.65$ if $P = 90$ (percent)	$.75 - 1.96(.10)$ to $.7 + 1.96(.10) = .55$ to $.95$
$z = 1.96$ if $P = 95$ (percent)	

Note. In Figure 1.1, $z = 2$ is used as an approximation to $z = 1.96$ when $p = 95\%$.

universe scores--at least those examinees about whom important decisions are to be made. If nothing else, such intervals are usually very revealing indicators of the amount of measurement error possibly involved in using $\bar{x}$ as if it were π. If an investigator feels that a specific interval is too broad for a specific decision, then the investigator might consider retesting the examinee.

Suppose, for example, that an examinee got 8 out of 10 items correct, initially, with a 67% credibility interval for π extending from .67 to 90. If the examinee were retested and got 10 out of 10 items correct, then for the combined tests $n = 20$, $x = 18$, and a 67% credibility interval extends from .82 to 95. This latter interval is considerably narrower than the former one; and, of course, the additional information supplied by the retest suggests that the examinee's universe score is probably higher than originally expected.

6. Group-Based Coefficients of Agreement and Measures of Error

Section 5 considered errors of measurement and errors of classification based on an individual examinee's score on a test. This section, considers issues involving group performance on a test. Specifically, the principal statistics to be discussed are indicated in Table 6.1.

The statistics $1 - p_o$ and $\sigma^2(\Delta)$ in Table 6.1 are closely related to errors of classification and errors of measurement, respectively. Specifically, $1 - p_o$ can be interpreted as the probability of an inconsistent decision; and $\sigma^2(\Delta)$ can be interpreted as the average value of the squared errors of measurement for examinees. As such, these statistics provide information about errors for a group of examinees, as opposed to an individual examinee.

The other statistics in Table 6.1 are called agreement coefficients in this handbook. Each of them has a value somewhere between 0 and 1, with higher values indicating greater degrees of agreement than lower values. The notion of "agreement" reflected by these coefficients involves considering what would happen (hypothetically) if examinees were administered many domain-referenced tests, with each test consisting of a different sample of n items from the universe. For a given test length (n), a high value for an agreement coefficient suggests that there would be a high degree of consistency in certain scores on these different tests. For example, if we knew that most persons classified as masters on one test would be classified as masters on most other tests, too, then one type of agreement would be relatively high. Although the above conceptual explanation of agreement coefficients rests on considering

Table 6.1

Loss Functions, Agreement Coefficients, and Errors

Based on Group Performance on a Test

Type of Loss	Agreement Coefficients Not Corrected For Chance	 Corrected For Chance	Errors
Threshold	p_o	Kappa	$1 - p_o$
Squared Error	$\Phi(c_o)$	Φ	$\sigma^2(\Delta)$

multiple tests, in practice these coefficients can be *estimated* using a single test, only; and in this handbook such single-test estimates are the only ones given detailed consideration.

The statistics in Table 6.1 can be classified into two categories based on the type of loss function involved in defining them. These two loss functions are called "threshold" loss and "squared error" loss. The subject of loss functions, *per se*, is a highly technical consideration that will not be treated in great detail here. For present purposes, it is sufficient to know that (a) a threshold loss function involves consideration of errors of classification, assumes that all false positive errors are equally serious, and assumes that all false negative errors are equally serious; and (b) a squared error loss function in domain-referenced testing involves consideration of errors of measurement and assumes that the seriousness of an error depends on (among other things) the squared distance between an examinee's observed and universe scores. Later, more will be said about these two loss functions; for now the reader should simply recognize that these two loss functions involve different approaches to addressing similar types of issues.

To develop some further understanding of the statistics in Table 6.1, suppose that test scores were available for a group of examinees on two forms of a domain-referenced test. Under this circumstance, the threshold loss coefficient denoted p_o in Table 6.1 would be

$$p_o = \left[\begin{array}{l}\text{Proportion of examinees classified as}\\ \text{masters on \underline{both} forms}\end{array}\right] + \left[\begin{array}{l}\text{Proportion of examinees classified}\\ \text{as non-masters on \underline{both} forms}\end{array}\right]$$

The coefficient p_o is, in effect, the proportion of examinees *consistently* classified into the same category (mastery or non-mastery) on the two tests.

It follows from the above paragraph that $1 - p_o$ is the proportion of examinees who are inconsistently classified on the two tests (i.e., classified as a master on one form and a non-master on the other). This proportion of inconsistent classifications is a group-based measure of error in a threshold loss sense, when scores on two tests are available.

The threshold loss coefficient p_o is not corrected for the expected "chance" agreement if all examinees were randomly assigned to a mastery or non-mastery status on each of the forms. The threshold-loss coefficient corrected for such chance agreement is called Kappa, which is defined as:

$$\text{Kappa} = (p_o - p_c)/(1 - p_c),$$

where p_c is chance agreement. In a sense, Kappa is a "pure" measure of agreement attributable to the testing procedure, under threshold loss assumptions.

The reader needs to be cautioned not to take the above "two-test" analogy too literally. It is offered simply as an aid in thinking about these statistics. Again, in this section the procedures treated involve a single administration of a single form of a domain-referenced test.

As noted in Table 6.1, corresponding to each of these three threshold loss statistics there is a statistic for squared error loss. For example, $\sigma^2(\Delta)$ is the average squared error of measurement for the population of examinees, and the two agreement coefficients for squared error loss involve $\sigma^2(\Delta)$. These squared error loss statistics provide a different perspective on agreement (and disagreement).

Throughout this section all reference to a cutting score, π_o , is replaced by consideration of $c_o = x_o/n$, the advancement score in terms of proportion of items correct. That is, in considering both squared error loss and threshold loss, c_o is sometimes used when it might be argued that π_o should be involved. To do so, however, would necessitate considerable complexities, no matter what loss function is involved.

Finally, it should be noted that some persons refer to the agreement coefficients discussed in this section as "reliability" coefficients. The word "reliability" is not used here principally to avoid unwarranted associations between the coefficients in Table 6.1 and classical reliability coefficients for norm-referenced tests. Given this caveat, however, much of this section treats issues traditionally associated with measurement consistency, or "reliability" considerations. (Also, in a sense mentioned later, these issues have validity connotations for domain-referenced interpretations.)

Squared Error Loss

Squared error loss statistics are conceptually more involved than their threshold loss counterparts. Here, however, intital consideration is given to squared error loss statistics because there are certain computational conveniences in proceeding in this order.

Suppose that an $n = 10$ item test were adminsitered to $k = 25$ examinees; and suppose that after the items were scored, the resulting data matrix was that given in Table 6.2. An entry in this data matrix is denoted x_{pi} , the score (0 = incorrect, 1 = correct) for examinee $\underline{p}$ on item $\underline{i}$.

Table 6.2

Group Performance on a Test:

A Synthetic Data Set with Sample Statistics

	Item										
Person	1	2	3	4	5	6	7	8	9	10	$\bar{x}_p$
1	1	1	1	1	1	1	1	1	1	1	1.0
2	1	1	1	1	1	1	1	1	1	1	1.0
3	1	1	1	1	1	1	1	1	1	1	1.0
4	1	1	1	1	1	1	1	1	1	1	1.0
5	1	1	1	1	1	1	1	1	1	1	1.0
6	1	1	1	1	1	1	1	1	1	1	1.0
7	1	1	1	1	1	1	1	1	1	1	1.0
8	1	1	0	1	1	1	1	1	1	1	.9
9	1	1	1	1	1	1	1	0	1	1	.9
10	1	1	1	0	1	1	1	1	1	1	.9
11	1	1	1	0	1	1	1	1	1	1	.9
12	1	1	1	1	1	1	1	1	1	0	.9
13	1	1	1	1	1	1	1	1	1	0	.9
14	1	0	1	1	1	1	0	1	1	1	.8
15	1	1	1	1	0	1	1	1	0	1	.8
16	1	1	1	1	0	1	1	0	1	1	.8
17	1	1	1	1	1	1	1	0	0	1	.8
18	1	1	1	1	1	1	1	1	0	0	.8
19	1	1	1	1	1	1	1	0	0	1	.8
20	0	0	1	1	1	0	1	1	1	1	.7
21	1	0	1	0	0	1	0	1	1	1	.6
22	1	0	1	1	1	1	0	1	0	0	.6
23	0	1	1	1	0	0	1	1	0	1	.6
24	1	0	1	1	1	1	0	0	0	0	.5
25	0	0	1	1	1	0	0	1	0	0	.4
$\bar{x}_i$	.88	.76	.96	.88	.84	.88	.80	.80	.68	.76	$\bar{x} = .824$
					$s^2(\bar{x}_i) = .0058$						$s^2(\bar{x}_p) = .0282$
					$s(\bar{x}_i) = .076$						$s(\bar{x}_p) = .168$

Other statistics reported in Table 6.2 are as follows:

(a) $\bar{x}_p$ is the proportion of items that examinee p got correct;

(b) $s^2(\bar{x}_p)$ and $s(\bar{x}_p)$ are the variance and standard deviation, respectively, of the scores $\bar{x}_p$;

(c) $\bar{x}_i$ is the proportion of persons who got item i correct--i.e., the item difficulty level discussed in Section 2;

(d) $s^2(\bar{x}_i)$ and $s(\bar{x}_i)$ are the variance and standard deviation, respectively, of the item difficulty levels; and

(e) $\bar{x}$ is the mean proportion of items correct for persons, or, equivalently, the mean difficulty level for items.

Using these sample statistics, Table 6.3 provides formulas, with illustrative computations, for estimating agreement coefficients and other quantities of interest involving squared error loss. (These formulas are used here because they are as computationally simple to use as any that can be derived; however, other more computationally difficult formulas would be better in terms of revealing certain underlying theoretical issues.)

Universe score variance. It has been emphasized repeatedly in previous sections that an examinee's observed score, $\bar{x}_p$, is not necessarily equal to his/her universe score, π_p. It follows that the variance of examinees' observed scores, $s^2(\bar{x}_p)$, is not necessarily equal to the variance of examinees' universe scores, $\sigma^2(\pi_p)$, which is abbreviated $\sigma^2(\pi)$ in Table 6.3. Actually, $\sigma^2(\pi)$ is almost always less than

Table 6.3

Equations and Illustrative Computations for Variances and Agreement Coefficients with Squared Error Loss

Equations	Computations Using Data in Table 6.2
Let k = number of examinees	$k = 25$
n = number of items	$n = 10$
$c_o = x_o/n$ = advancement score in terms of proportion of items correct	$c_o = 9/10 = .9$
Universe Score Variance	
(6.1) $\sigma^2(\pi) = \dfrac{k[ns^2(\bar{x}_p) + s^2(\bar{x}_i) - \bar{x}(1-\bar{x})]}{(n-1)(k-1)}$	$\sigma^2(\pi) = \dfrac{25[10(.0282) + .0058 - .824(1-.824)]}{(9)(24)}$ $= .0165 \qquad [\sigma(\pi) = \sqrt{.0165} = .129]$
Error Variance	
(6.2) $\sigma^2(\Delta) = \dfrac{\bar{x}(1-\bar{x}) - s^2(\bar{x}_p)}{n - 1}$	$\sigma^2(\Delta) = \dfrac{.824(1-.824) - .0282}{9}$ $= .0130 \qquad [\sigma(\Delta) = \sqrt{.0130} = .114]$
Agreement Coefficient <u>Not</u> Corrected for Chance	
(6.3) $\Phi(c_o) = \dfrac{s^2(\bar{x}_p) + (\bar{x} - c_o)^2 - \sigma^2(\Delta)}{s^2(\bar{x}_p) + (\bar{x} - c_o)^2}$	$\Phi(.9) = \dfrac{.0282 + (.824 - .9)^2 - .0130}{.0282 + (.824 - 9)^2}$ $= .617$
KR-21	
(6.4) $\text{KR-21} = \Phi(c_o = \bar{x}) = \dfrac{s^2(\bar{x}_p) - \sigma^2(\Delta)}{s^2(\bar{x}_p)}$	$\text{KR-21} = \dfrac{.0282 - .0130}{.0282}$ $= .539$
Agreement Coefficient Corrected for Chance	
(6.5) $\Phi = \dfrac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\Delta)}$	$\Phi = \dfrac{.0165}{.0165 + .0130}$ $= .559$

the observed score variance. This fact is not immediately evident from Equation 6.1 in Table 6.3; but the computation section of Table 6.3 shows that $\sigma^2(\pi) = .0165$, a value considerably smaller than $s^2(\bar{x}_p) = .0282$. Note that the square root of $\sigma^2(\pi)$ is simply the standard deviation of examinee universe scores, which is $\sigma^2(\pi) = .129$ for the synthetic data.

Error variance. Recall from Section 5 that error of measurement is defined as the difference between an examinee's observed and universe scores:

$$\Delta_p = \bar{x}_p - \pi_p \quad .$$

If we were to square these differences for all examinees, and then get the average of these squared differences, we would obtain $\sigma^2(\Delta)$. Of course, π_p is never known exactly, so neither is Δ_p; and, consequently, $\sigma^2(\Delta)$ cannot be obtained directly by averaging the squared values of Δ_p. However, one can estimate $\sigma^2(\Delta)$ using Equation 6.2 in Table 6.3, and the square root of this value is an estimate of the standard deviation of examinee errors of measurement. For the data in Table 6.2, Table 6.3 shows that $\sigma^2(\Delta) = .0130$ and $\sigma(\Delta) = .114$. It is not immediately evident from Table 6.3, but $\sigma^2(\Delta)$ depends upon the variance of item difficulty levels, among other things. In general, the smaller the variance of item difficulty levels, the smaller the value of $\sigma^2(\Delta)$.

Agreement coefficient not corrected for chance. The above discussion of universe score variance and error variance makes no reference to mastery/non-mastery decisions. When such decisions are to be made, the advancement score plays a role in the definition of an agreement coefficient not corrected for chance, although error variance is still $\sigma^2(\Delta)$. This agreement coefficient is defined as:

$$\Phi(c_o) = \frac{\sigma^2(\pi) + (\mu - c_o)^2}{\sigma^2(\pi) + (\mu - c_o)^2 + \sigma^2(\Delta)} ;$$

where $c_o = x_o/n$ is the advancement score in terms of proportion of items correct; and μ is the mean score over the universe of items and the population of persons. As such, μ has similarities with $\bar{x}$, but is not identical to it. The above definition is rather difficult to use directly to estimate $\Phi(c_o)$, so a simpler formula is provided by Equation 6.3 in Table 6.3.

Note that Equation 6.3 depends upon $(\bar{x} - c_o)^2$, the squared difference between $\bar{x}$ and the advancement score. For the synthetic data with $\bar{x} = .824$ and $c_o = 9/10 = .9$, Table 6.3 shows that $\Phi(.9) = .62$. One might ask, however, what would be the value of $\Phi(c_o)$ if $\bar{x}$ actually equaled c_o in Equation 6.3? The answer is provided by Equation 6.4, which is also identified as KR-21. As discussed later, KR-21 also plays an important role in estimating threshold loss agreement coefficients. For the synthetic data KR-21 = .54, and this is the smallest value that Equation 6.3 can have for these data--no matter what the advancement score actually is.

Agreement coefficient corrected for chance. The agreement coefficient corrected for chance, which is denoted Φ, is easily obtained using the values of $\sigma^2(\pi)$ and $\sigma^2(\Delta)$ in Equation 6.5 in Table 6.3. For the synthetic data, $\Phi = .56$, a value very close to KR-21 = .54. Indeed, Φ and KR-21 almost always have very similar values. This occurs principally because neither one of them depends on chance agreement, which is technically $(\mu - c_o)^2$ for squared error loss.

Interpreting agreement coefficients. Agreement coefficients (and their reliability counterparts) are discussed and used extensively in educational measurement--perhaps too extensively! However, they are frequently difficult to interpret correctly, no matter what loss function is involved. For this reason, whatever loss function is involved, the following characteristics of such coefficients should be kept in mind

(a) an agreement coefficient generally ranges from 0 to 1, but a value of ,say, .80 is not necessarily "twice as good" as a value of .40;

(b) when most examinees have observed scores close to the advancement score, an agreement coefficient not corrected for chance will be smaller than when most examinees have observed scores relatively far from the advancement score;

(c) an agreement coefficient will tend to be small whenever universe score variance is small or error variance is large (even if the coefficient is based on threshold loss);

(d) an agreement coefficient not corrected for chance reflects the quality (or consistency) of decisions made about examinees, whereas an agreement coefficient corrected for chance reflects the contribution of the test to the quality of such decisions. This is another perspective on the fact that a coefficient corrected for chance is smaller than its not-corrected-for-chance counterpart.

Threshold Loss

In the introduction to this section it was stated that a threshold loss function assumes that all false negative errors are equally serious, and all false positive errors are equally serious.

To clarify this point let us suppose that the test length is n = 10, and $c_o = \pi_o = .90$. Obviously, an examinee will not be advanced if he/she gets 0, 1, 2, . . ., 8 items correct. Now, it is almost certain that some of these examinees will be falsely classified as non-masters, because it is likely that some of these examinees have universe scores at or above .90. (Of course, one never knows which examinees are falsely declared to be non-masters). For threshold loss it is assumed that any such false negative error is as serious as any other such error, no matter what the examinee's unvierse score actually is; e.g., failing an examinee with a universe score of $\pi = .91$ is as serious an error as failing an examinee with a universe score of $\pi = 1.00$.

Also, the threshold loss function involves assuming that all false positive errors are equally serious. For the above example, this means that passing an examinee with a universe score of, say, $\pi = .40$ is as serious an error as passing an examinee with a universe score of, say, $\pi = .70$.

It should be noted, however, that the threshold loss function does not involve assuming that false positive errors are as serious as false negative errors. That issue is a question of loss ratio--a subject treated in Section 4.

Table 6.4 describes and illustrates the steps required to obtain the threshold loss coefficients p_o (not corrected for chance) and Kappa (corrected for chance).

Step 1 simply involves recording results already obtained in Tables 6.2 and 6.3 for the synthetic data.

Step 2 involves computing a z-score based on the advancement score, c_o. For these data $z = .45$ which means that the mean, $\bar{x}$, is 45/100th's of a standard deviation $[s(\bar{x}_p) = .168]$ above the advancement score.

Step 3 involves determining what proportion of examinees would have z-scores below $z = .45$ if examinee scores were normally distributed. To obtain this result, Table A.3 in Appendix A is required. For the synthetic data, this proportion is $p_z = .67$.

Step 4 involves determining the proportion of examinees who would have z-scores below $z = .45$ on each of two (hypothetical) n-item tests, if examinee scores were normally distributed on both tests. For the synthetic data $p_{zz} = .53$. This step makes use of KR-21; and p_{zz} will

Table 6.4

A Procedure with Illustrative Computations for Estimating Agreement Coefficients and Expected Proportion of Inconsistent Decisions with Threshold Loss

	Procedure	Example Using Data in Table 6.2
Step 1:	Specify $c_o = x_o/n$ and calculate $\bar{x}$, $s(\bar{x}_p)$, and KR-21 (see Equation 6.4 in Table 6.3)	$c_o = 9/10 = .9$ $\bar{x} = .824 \qquad s(\bar{x}_p) = .168$ $\text{KR-21} = .539$
Step 2:	Compute the z-score corresponding to c_o $z = (c_o - \bar{x})/s(\bar{x}_p)$	$z = (.9 - .824)/.168 = .452$
Step 3:	Using the last column in Table A.3, locate the row having the closest value to the z-score in Step 2. Record p_z--the entry to the left of this z-score (under the column headed 1.00)	$p_z = .674$
Step 4:	Find the column in Table A.3 having the closest value to KR-21 in Step 1. Record p_{zz}--the entry in this column for the row located in Step 3.	Using the column headed KR-21 = .55, $p_{zz} = .533$
Step 5:	Compute p_o and kappa $p_o = 1 - 2(p_z - p_{zz})$ $\text{Kappa} = \dfrac{p_{zz} - p_z^2}{p_z - p_z^2}$	$p_o = 1 - 2(.674 - .533) = .72$ $\text{Kappa} = \dfrac{.533 - (.674)^2}{.674 - (.674)^2} = .36$
Step 6:	Compute the expected proportion of inconsistent decisions $1 - p_o$	$1 - p_o = .28$

always be less than p_z , unless KR-21 actually equals one (a highly unlikely occurrence).

Step 5 provides formulas for estimating p_o and Kappa using p_z and p_{zz}. For the synthetic data p_o = .72, and Kappa = .36. Again, Kappa is smaller than p_o because p_o reflects the proportion of examinees consistently classified, while Kappa reflects the proportion of examinees consistently classified over and beyond the proportion that would probably be classified consistently by chance. [The proportion probably classified consistently by chance is $1 - 2 p_z (1 - p_z)$, which is .54 for the synthetic data.]

Finally, Step 6 in Table 6.4 provides an estimate of the proportion of examinees who are inconsistently classified, i.e., the proportion of errors involved in the decision-making process, in the sense of threshold loss errors. For the synthetic data, this proportion is .28.

The procedure for estimating p_o and Kappa in Table 6.4 is based on the assumption that examinee universe scores are normally distributed. In many domain-referenced testing contexts this assumption is probably not true; but in most cases it is unlikely that violations of this assumption will cause p_o and Kappa to be poorly estimated.

It is important to note that the statistics discussed above refer to a group of examinees--not to individual examiness. None of these statistics specify which examinees are consistently or inconsistently classified.

Also, for a different group of examinees, and/or a different sample of items, the results would almost certainly differ. A similar statement applies to the statistics for squared error loss in Table 6.3. Such differences do not invalidate the statistics discussed above; rather, such differences result because what we are really doing is estimating quantities (called parameters) that we cannot observe directly.

Test Length

Recall that a domain-referenced test is viewed as a sample of items from a larger universe of items constructed to measure the content under consideration. Also recall that the examinee scores one would ideally like to know are the examinee universe scores--i.e., examinee scores on the universe of items. These ideal scores can never be obtained; but, in general, longer tests involve less error and provide better estimates of examinee universe scores.

Therefore, one obvious question is, "How long should a test be?" There can be no universal statistical answer to this question, because any specific attempt to answer it eventually involves answering at least one other question--namely, "How much error is one willing to tolerate?" Clearly, the answer to this latter question necessitates subjective judgment by a responsible person who is well-aware of all aspects of the testing environment and the decisions to be made. Even so, statistics can help in making informed subjective judgments about test length.

In particular two such statistics can be helpful: (a) $\sigma(\Delta)$, the standard deviation of errors of measurement; and (b) $1 - p_o$, the proportion of examinees inconsistently classified. Table 6.5 shows how these two statistics can be estimated for a hypothetical test of length $\underline{n}'$. Actually, only Equation 6.6 in Table 6.5 is required to estimate error variance and its standard deviation; the other equations and steps are required to obtain the proportion of examinees inconsistently classified.

Note that in Table 6.5 statistics for a test of length $\underline{n}'$ are identified with a prime to distinguish them from the corresponding statistics for the available n-item test. This distinction is dropped in Table 6.6 which summarizes results for test lengths of n = 10, 15, and 20. (The first row of Table 6.6 simply duplicates results already reported in Tables 6.3 and 6.4 for the 10-item test.) From Table 6.6 it is clear that, as test length increases, both $\sigma(\Delta)$ and $1 - p_o$ decrease, but not very rapidly. In interpreting $\sigma(\Delta)$ it is useful to keep in mind that it can be no larger than 0.25 when each observed item score takes on one of two possible values, as is the case for the the synthetic data in Table 6.2.

The values of $\sigma(\Delta)$ and $1 - p_o$ reported in Table 6.6 are based upon synthetic data, but similar results can easily occur with real data. Furthermore, the values of $\sigma(\Delta)$ and $1 - p_o$ reported in Table 6.6 would probably be judged rather large in most real contexts. Of course, these values can be reduced by increasing test length beyond 20 items.

Table 6.5

Equations and Procedures for Examining Effect of Changes in Test Length

Equations and Procedures	Example Using Results in Tables 6.3 and 6.4
Given n, $\bar{x}$, $\sigma^2(\Delta)$, KR-21, an indifference zone, and a loss ratio for an available test.	$n = 10$ $\bar{x} = .824$ $\sigma^2(\Delta) = .0130$ KR-21 = .539 Indifference Zone = (.75,.85) Loss Ratio = 2
Let: n' = number of items in hypothetical test	$n' = 20$
(6.6) $\sigma^2(\Delta') = n\,\sigma^2(\Delta)/n'$	$\sigma^2(\Delta') = 10(.0130)/20 = .0065$ $[\sigma(\Delta') = .081]$
(6.7) $(\text{KR-21})' = \dfrac{n'(\text{KR-21})}{n + (n'-n)(\text{KR-21})}$	$(\text{KR-21})' = \dfrac{20(.539)}{10 + (20-10)(.539)} = .701$
(6.8) $s^2(\bar{x}'_p) = \dfrac{\sigma^2(\Delta')}{1 - (\text{KR-21})'}$	$s^2(\bar{x}'_p) = \dfrac{.0065}{1 - .701} = .0217$ $[s(\bar{x}'_p) = .147]$
p_o for a Test of Length n' (i.e., p'_o):	
(a) Using Table A.1 , determine x'_o for a test of length n', given the indifference zone and loss ratio for the available test	(a) For a 20-item test, with an indifference zone of (.75, .85) and a loss ratio of 2, Table A.2 indicates that $x'_o = 17$
(b) $c'_o = x'_o/n'$	(b) $c'_o = 17/20 = .850$
(c) Use c'_o , $(\text{KR-21})'$, and $s(\bar{x}'_p)$, as calculated above (and $\bar{x}$ for the available test) in the procedure in Table 6.4	(c) $z' = (.850 - .824)/.147 = .177$ $p'_z = .579$ and $p'_{zz} = .455$ Therefore, $p'_o = .75$.

Note: All statistics for a test of length n' are identified with a prime (′) to distinguish them from the corresponding statistics for the available n-item test.

Table 6.6

Illustrative Results for Changes in Test Length Using the Synthetic Data Example

n	$\sigma(\Delta)$	KR-21	$1-p_o$
10	.11	.54	.28
15	.09	.64	.26
20	.08	.70	.25

In beginning the above discussion of test length, it was pointed out that data, per se, cannot specify what the test length should be, but data can help in making an informed, but still subjective, judgment about test length. In this regard, $\sigma(\Delta)$ and $1 - p_o$ are helpful; but it must be recognized that these two statistics provide different types of information, and perhaps not equally useful information in a particular context. In the extreme, if an investigator were interested only in minimizing classification errors, then $\sigma(\Delta)$ would provide irrelevant information; and, conversely, if an investigator were interested only in measurement error, then $1 - p_o$ would provide irrelevant information.

The perspective taken above is that in most realistic settings, both types of error are likely to be of interest; and, therefore, consideration has been given to both. Only in a specific context can a judgment be made concerning which statistic is more appropriate in considerations regarding test length. As discussed below, a similar argument applies to agreement coefficients.

Other Considerations

Throughout this section, squared error loss and threshold loss statistics have been treated in parallel. If, in a given context, an investigator has an unambiguous basis for choosing one loss function over the other, then, of course, statistics involving the other loss function become irrelevant. However, in many situations, choice of a loss function may not be a completely unambiguous decision and, indeed, it may be that neither loss function is ideal. In such situations, one approach is to examine statistics for both loss functions, keeping in mind the different assumptions involved. In doing so, there is some potential for confusion, but a theoretically better approach would involve complexities far beyond the intended scope of this handbook.

In this regard, it should be kept in mind that it is not always the case that a test is used to make a single type of decision. For example, it could well be that a given test is sometimes used to make mastery/non-mastery types of decisions assuming threshold loss; and, at other times, the test is used simply to estimate examinee universe scores assuming squared error loss. For such a test, both loss functions are appropriate depending upon the use of the test. Indeed, in choosing a loss function, the question of importance is not what constitutes the test, but rather what constitutes the assumptions about the decisions to be made using the test.

Sometimes a domain-referenced test is used solely for the purpose of estimating examinee universe scores, without any consideration of a cutting score. In such situations (assuming that squared error loss is relevant), $\sigma(\Delta)$ is still appropriate, as is the index Φ given by

Equation 6.5 in Table 6.3. In this sense, Φ may be viewed as a general-purpose agreement coefficient, or index of dependability, for a domain-referenced test. Note that when a domain-referenced test is used solely to estimate examinee universe scores, threshold loss statistics like those treated above are meaningless.

In the introduction to this section, reference was made to the fact that the agreement coefficients discussed above are sometimes called reliability coefficients. Actually, these agreement coefficients carry with them a connotation of validity, too, in the sense that they involve consideration of the universe of items which is often the principal "criterion" of interest, or the only criterion available. Indeed, one perspective on measurement suggests that notions of reliability and validity can be blended together into a consideration of the extent to which observed scores are generalizable to universe scores. This perspective seems especially relevant for domain-referenced interpretations of test scores. In this sense, this section has considered issues relevant to both reliability and validity.

Appendix A

Tables

Table A.1 is based on the Fhanér-Wilcox-Huynh procedure referenced in Appendix B. This table was developed using the IMSL (1979) subroutine MDBETA.

The results reported in Table A.2 are based on the assumptions of binomial likelihood and a uniform beta prior (see Appendix B). The probabilities reported in Table A.2 were obtained using the IMSL (1979) subroutine MDBETA: and the credibility intervals were obtained using CADA [Isaacs and Novick, and Jackson (1974)], and some calculus.

Table A.3 was developed using the IMSL (1979) subroutines MDBNOR and MDNOR.

Table A.1

ADVANCEMENT SCORES FOR VARIOUS INDIFFERENCE ZONES ORDERED ACCORDING TO INTERVAL MID-POINTS FOR LOSS RATIOS OF 1, 2, AND 3

	MID-POINT = 0.65		MID-POINT = 0.70				MID-POINT = 0.75				
	0.625 TO 0.675	0.600 TO 0.700	0.675 TO 0.725	0.650 TO 0.750	0.625 TO 0.775	0.600 TO 0.800	0.725 TO 0.775	0.700 TO 0.800	0.675 TO 0.825	0.650 TO 0.850	
n	(0.050)	(0.100)	(0.050)	(0.100)	(0.150)	(0.200)	(0.050)	(0.100)	(0.150)	(0.200)	n
6	4/ 5/ 5	4/ 5/ 5	5/ 5/ 6	5/ 5/ 6	5/ 5/ 6	5/ 5/ 5	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	6
7	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	6/ 6/ 7	6/ 6/ 7	6/ 6/ 7	6/ 6/ 7	7
8	6/ 6/ 7	6/ 6/ 7	6/ 7/ 7	6/ 7/ 7	6/ 7/ 7	6/ 7/ 7	7/ 7/ 8	7/ 7/ 7	7/ 7/ 7	7/ 7/ 7	8
9	6/ 7/ 8	6/ 7/ 7	7/ 8/ 8	7/ 7/ 8	7/ 7/ 8	7/ 7/ 8	7/ 8/ 8	7/ 8/ 8	7/ 8/ 8	7/ 8/ 8	9
10	7/ 8/ 8	7/ 8/ 8	8/ 8/ 9	8/ 8/ 9	8/ 8/ 8	8/ 8/ 8	8/ 9/ 9	8/ 9/ 9	8/ 9/ 9	8/ 9/ 9	10
11	8/ 8/ 9	8/ 8/ 9	8/ 9/ 9	8/ 9/ 9	8/ 9/ 9	8/ 9/ 9	9/10/10	9/ 9/10	9/ 9/10	9/ 9/10	11
12	8/ 9/10	8/ 9/ 9	9/10/10	9/10/10	9/10/10	9/10/10	10/10/11	10/10/11	10/10/10	10/10/10	12
13	9/10/10	9/10/10	10/10/11	10/10/11	10/10/11	10/10/11	10/11/11	10/11/11	10/11/11	10/11/11	13
14	10/10/11	10/10/11	10/11/12	10/11/11	10/11/11	10/11/11	11/12/12	11/12/12	11/12/12	11/12/12	14
15	10/11/12	10/11/11	11/12/12	11/12/12	11/12/12	11/12/12	12/13/13	12/12/13	12/12/13	12/12/13	15
16	11/12/12	11/12/12	12/13/13	12/12/13	12/12/13	12/12/13	13/13/14	13/13/14	13/13/14	13/13/13	16
17	12/12/13	12/12/13	12/13/14	12/13/14	13/13/13	13/13/13	13/14/15	13/14/14	13/14/14	13/14/14	17
18	12/13/14	12/13/13	13/14/14	13/14/14	13/14/14	13/14/14	14/15/15	14/15/15	14/15/15	14/15/15	18
19	13/14/14	13/14/14	14/15/15	14/15/15	14/15/15	14/14/15	15/16/16	15/16/16	15/15/16	15/15/16	19
20	14/14/15	14/14/15	15/15/16	15/15/16	15/15/16	15/15/16	16/16/17	16/16/17	16/16/17	16/16/17	20
21	14/15/16	14/15/15	15/16/17	15/16/16	15/16/16	15/16/16	16/17/18	16/17/17	16/17/17	17/17/17	21
22	15/16/16	15/16/16	16/17/17	16/17/17	16/17/17	16/17/17	17/18/18	17/18/18	17/18/18	17/18/18	22
23	16/16/17	16/16/17	17/18/18	17/17/18	17/17/18	17/17/18	18/19/19	18/19/19	18/19/19	18/19/19	23
24	16/17/18	16/17/17	17/18/19	17/18/19	17/18/18	18/18/18	19/19/20	19/19/20	19/19/20	19/19/20	24
25	17/18/18	17/18/18	18/19/20	18/19/19	18/19/19	18/19/19	19/20/21	19/20/21	19/20/20	20/20/20	25
26	17/18/19	17/18/19	19/20/20	19/20/20	19/19/20	19/19/20	20/21/21	20/21/21	20/21/21	20/21/21	26
27	18/19/20	18/19/19	19/20/21	20/20/21	20/20/21	20/20/20	21/22/22	21/22/22	21/22/22	21/22/22	27
28	19/20/20	19/20/20	20/21/22	20/21/21	20/21/21	20/21/21	22/23/23	22/22/23	22/22/23	22/22/23	28
29	19/20/21	19/20/21	21/22/22	21/22/22	21/22/22	21/22/22	22/23/24	22/23/24	22/23/23	23/23/23	29
30	20/21/22	20/21/21	22/23/23	22/22/23	22/22/23	22/22/23	23/24/25	23/24/24	23/24/24	23/24/24	30

NOTE. THE WIDTH OF EACH INTERVAL IS INDICATED IN PARENTHESES BELOW THE LIMITS OF THE INTERVAL. ENTRIES IN THE TABLE ARE ADVANCEMENT SCORES FOR LOSS RATIOS OF 1, 2, AND 3, RESPECTIVELY. FOR EXAMPLE, A LOSS RATIO OF 2 IS APPROPRIATE IF FALSE POSITIVE ERRORS ARE TWICE AS SERIOUS AS FALSE NEGATIVE ERRORS.

Table A.1 (Continued)

ADVANCEMENT SCORES FOR VARIOUS INDIFFERENCE ZONES
ORDERED ACCORDING TO INTERVAL MID-POINTS
FOR LOSS RATIOS OF 1, 2, AND 3

	MID-POINT = 0.80				MID-POINT = 0.85				MID-POINT = 0.90		
	0.775 TO 0.825	0.750 TO 0.850	0.725 TO 0.875	0.700 TO 0.900	0.825 TO 0.875	0.800 TO 0.900	0.775 TO 0.925	0.750 TO 0.950	0.875 TO 0.925	0.850 TO 0.950	
n	(0.050)	(0.100)	(0.150)	(0.200)	(0.050)	(0.100)	(0.150)	(0.200)	(0.050)	(0.100)	n
6	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	5/ 6/ 6	6/ 6/ 6	6/ 6/ 6	6/ 6/ 6	6/ 6/ 6	6/ 6/ 6	6/ 6/ 6	6
7	6/ 7/ 7	6/ 7/ 7	6/ 7/ 7	6/ 7/ 7	7/ 7/ 7	7/ 7/ 7	7/ 7/ 7	7/ 7/ 7	7/ 7/ 7	7/ 7/ 7	7
8	7/ 8/ 8	7/ 8/ 8	7/ 7/ 8	7/ 7/ 8	7/ 8/ 8	7/ 8/ 8	7/ 8/ 8	8/ 8/ 8	8/ 8/ 8	8/ 8/ 8	8
9	8/ 8/ 9	8/ 8/ 9	8/ 8/ 9	8/ 8/ 8	8/ 9/ 9	8/ 9/ 9	8/ 9/ 9	8/ 9/ 9	9/ 9/ 9	9/ 9/ 9	9
10	9/ 9/10	9/ 9/ 9	9/ 9/ 9	9/ 9/ 9	9/10/10	9/10/10	9/10/10	9/10/10	10/10/10	10/10/10	10
11	9/10/10	9/10/10	9/10/10	10/10/10	10/11/11	10/10/11	10/10/11	10/10/11	11/11/11	11/11/11	11
12	10/11/11	10/11/11	10/11/11	10/11/11	11/11/12	11/11/12	11/11/12	11/11/12	11/12/12	11/12/12	12
13	11/12/12	11/12/12	11/12/12	11/12/12	12/12/13	12/12/12	12/12/12	12/12/12	12/13/13	12/13/13	13
14	12/12/13	12/12/13	12/12/13	12/12/13	13/13/13	13/13/13	13/13/13	13/13/13	13/14/14	13/14/14	14
15	13/13/14	13/13/14	13/13/13	13/13/13	13/14/14	13/14/14	13/14/14	14/14/14	14/15/15	14/15/15	15
16	13/14/15	13/14/14	13/14/14	14/14/14	14/15/15	14/15/15	14/15/15	15/15/15	15/16/16	15/16/16	16
17	14/15/15	14/15/15	14/15/15	14/15/15	15/16/16	15/16/16	15/16/16	15/16/16	16/16/17	16/16/17	17
18	15/16/16	15/16/16	15/16/16	15/16/16	16/17/17	16/16/17	16/16/17	16/17/17	17/17/18	17/17/18	18
19	16/17/17	16/16/17	16/16/17	16/16/17	17/17/18	17/17/18	17/17/18	17/17/18	18/18/19	18/18/18	19
20	17/17/18	17/17/18	17/17/18	17/17/17	18/18/19	18/18/19	18/18/18	18/18/19	19/19/20	19/19/19	20
21	17/18/19	17/18/18	18/18/18	18/18/18	18/19/20	19/19/19	19/19/19	19/19/19	20/20/20	20/20/20	21
22	18/19/19	18/19/19	18/19/19	18/19/19	19/20/20	19/20/20	20/20/20	20/20/20	20/21/21	21/21/21	22
23	19/20/20	19/20/20	19/20/20	19/20/20	20/21/21	20/21/21	20/21/21	21/21/21	21/22/22	21/22/22	23
24	20/21/21	20/20/21	20/20/21	20/21/21	21/22/22	21/22/22	21/22/22	21/22/22	22/23/23	22/23/23	24
25	21/21/22	21/21/22	21/21/22	21/21/22	22/23/23	22/23/23	22/23/23	22/23/23	23/24/24	23/24/24	25
26	21/22/23	21/22/22	22/22/22	22/22/22	23/23/24	23/23/24	23/23/24	23/24/24	24/25/25	24/25/25	26
27	22/23/24	22/23/23	22/23/23	23/23/23	24/24/25	24/24/25	24/24/25	24/24/25	25/26/26	25/26/26	27
28	23/24/24	23/24/24	23/24/24	23/24/24	24/25/26	25/25/25	25/25/25	25/25/26	26/26/27	26/26/27	28
29	24/25/25	24/25/25	24/25/25	24/25/25	25/26/26	25/26/26	26/26/26	26/26/26	27/27/28	27/27/28	29
30	25/25/26	25/25/26	25/25/26	25/25/26	26/27/27	26/27/27	26/27/27	27/27/27	28/28/29	28/28/29	30

NOTE. THE WIDTH OF EACH INTERVAL IS INDICATED IN PARENTHESES BELOW THE LIMITS OF THE INTERVAL. ENTRIES IN THE TABLE ARE ADVANCEMENT SCORES FOR LOSS RATIOS OF 1, 2, AND 3, RESPECTIVELY. FOR EXAMPLE, A LOSS RATIO OF 2 IS APPROPRIATE IF FALSE POSITIVE ERRORS ARE TWICE AS SERIOUS AS FALSE NEGATIVE ERRORS.

Table A.2

Inferences about Universe Score Given n and x for an Examinee

			Probability that π is at or above								Credibility Intervals		
n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
5	3	.60	46	35	26	17	10	5	2	0	(.41, .77)	(.35, .82)	(.29, .86)
5	4	.80	77	68	58	47	34	22	11	3	(.62, .92)	(.55, .95)	(.48, .97)
5	5	1.00	95	92	88	82	74	62	47	26	(.83, 1.00)	(.76, 1.00)	(.68, 1.00)
6	3	.50	29	20	13	7	3	1	0	0	(.33, .67)	(.28, .72)	(.23, .77)
6	4	.67	58	47	35	24	15	7	3	0	(.49, .82)	(.43, .85)	(.37, .89)
6	5	.83	84	77	67	56	42	28	15	4	(.67, .94)	(.61, .96)	(.54, .98)
6	6	1.00	97	95	92	87	79	68	52	30	(.85, 1.00)	(.79, 1.00)	(.72, 1.00)
7	4	.57	41	29	19	11	6	2	1	0	(.41, .73)	(.36, .77)	(.30, .82)
7	5	.71	68	57	45	32	20	11	4	1	(.55, .84)	(.50, .88)	(.43, .91)
7	6	.86	89	83	74	63	50	34	19	6	(.71, .95)	(.65, .97)	(.58, .98)
7	7	1.00	98	97	94	90	83	73	57	34	(.87, 1.00)	(.82, 1.00)	(.75, 1.00)
8	4	.50	27	17	10	5	2	1	0	0	(.35, .65)	(.30, .70)	(.25, .75)
8	5	.63	52	37	27	17	9	3	1	0	(.42, .77)	(.42, .80)	(.36, .85)
8	6	.75	77	66	54	40	26	14	5	1	(.60, .86)	(.55, .90)	(.48, .93)
8	7	.88	93	88	80	70	56	40	23	7	(.74, .96)	(.69, .97)	(.62, .98)
8	8	1.00	99	98	96	92	87	77	61	37	(.89, 1.00)	(.84, 1.00)	(.77, 1.00)
9	5	.56	37	25	15	8	3	1	0	0	(.41, .70)	(.37, .74)	(.31, .78)
9	6	.67	62	49	35	22	12	5	1	0	(.52, .79)	(.47, .83)	(.41, .87)
9	7	.78	83	74	62	47	32	18	7	1	(.64, .88)	(.59, .91)	(.53, .94)
9	8	.89	95	91	85	76	62	46	26	9	(.77, .96)	(.77, .98)	(.66, .99)
9	9	1.00	99	99	97	94	89	80	65	40	(.90, 1.00)	(.85, 1.00)	(.79, 1.00)
10	5	.50	25	15	8	3	1	0	0	0	(.36, .64)	(.32, .68)	(.27, .73)
10	6	.60	47	33	21	11	5	2	0	0	(.46, .73)	(.41, .77)	(.36, .81)
10	7	.70	70	57	43	29	16	7	2	0	(.56, .82)	(.51, .85)	(.46, .88)
10	8	.80	88	80	69	54	38	22	9	2	(.67, .90)	(.62, .92)	(.56, .94)
10	9	.90	97	94	89	80	68	51	30	10	(.79, .97)	(.74, .98)	(.69, .98)
10	10	1.00	100	99	98	96	91	83	69	43	(.90, 1.00)	(.86, 1.00)	(.81, 1.00)

Table A.2 (Continued)

Inferences about Universe Score Given n and x for an Examinee

			Probability that π is at or above								Credibility Intervals		
n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
11	6	.55	33	21	12	5	2	0	0	0	(.41, .68)	(.37, .72)	(.35, .76)
11	7	.64	56	42	28	16	7	2	0	0	(.50, .76)	(.45, .75)	(.40, .83)
11	8	.73	77	65	51	35	21	9	3	0	(.60, .84)	(.55, .87)	(.50, .90)
11	9	.82	92	85	75	61	44	26	11	2	(.70, .90)	(.66, .93)	(.60, .95)
11	10	.91	98	96	91	84	73	56	34	12	(.80, .97)	(.76, .98)	(.71, .94)
11	11	1.00	100	99	99	97	93	86	72	46	(.91, 1.00)	(.87, 1.00)	(.83, 1.00)
12	6	.50	23	13	6	2	1	0	0	0	(.37, .63)	(.33, .67)	(.29, .71)
12	7	.58	43	28	17	8	3	1	0	0	(.45, .71)	(.41, .74)	(.36, .78)
12	8	.67	65	50	35	21	10	3	1	0	(.54, .78)	(.49, .81)	(.44, .85)
12	9	.75	83	72	58	42	25	12	3	0	(.63, .85)	(.58, .88)	(.53, .91)
12	10	.83	94	89	80	67	50	31	13	2	(.72, .92)	(.68, .94)	(.62, .96)
12	11	.92	99	97	94	87	77	60	38	14	(.82, .97)	(.78, .98)	(.73, .99)
12	12	1.00	100	100	99	98	95	88	75	49	(.92, 1.00)	(.88, 1.00)	(.84, 1.00)
13	7	.54	31	18	9	4	1	0	0	0	(.41, .66)	(.38, .70)	(.33, .74)
13	8	.62	51	36	22	11	4	1	0	0	(.49, .73)	(.45, .77)	(.40, .80)
13	9	.69	72	58	42	26	13	5	1	0	(.57, .80)	(.53, .83)	(.47, .86)
13	10	.77	88	78	64	48	30	15	4	0	(.65, .86)	(.61, .89)	(.56, .91)
13	11	.85	96	92	84	72	55	35	16	3	(.74, .92)	(.70, .94)	(.65, .96)
13	12	.92	99	98	95	90	80	64	42	15	(.83, .97)	(.79, .98)	(.75, .99)
13	13	1.00	100	100	99	98	96	90	77	51	(.92, 1.00)	(.89, 1.00)	(.85, 1.00)
14	7	.50	21	11	5	2	0	0	0	0	(.38, .62)	(.34, .66)	(.30, .70)
14	8	.57	39	25	13	6	2	0	0	0	(.45, .68)	(.41, .72)	(.37, .76)
14	9	.64	60	44	28	15	6	2	0	0	(.52, .75)	(.48, .79)	(.43, .82)
14	10	.71	78	65	48	31	16	6	1	0	(.60, .82)	(.55, .84)	(.51, .87)
14	11	.79	91	83	70	54	35	18	6	1	(.67, .87)	(.63, .90)	(.59, .92)
14	12	.86	97	94	87	76	60	40	18	4	(.76, .93)	(.72, .95)	(.67, .96)
14	13	.93	99	99	96	92	83	68	45	17	(.84, .98)	(.81, .99)	(.76, .99)
14	14	1.00	100	100	100	99	96	91	79	54	(.93, 1.00)	(.90, 1.00)	(.86, 1.00)

Table A.2 (Continued)

Inferences about Universe Score Given
n and x for an Examinee

			Probability that π is at or above								Credibility Intervals		
n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
15	8	.53	28	16	7	3	1	0	0	0	(.42, .65)	(.38, .66)	(.34, .72)
15	9	.60	47	31	18	8	3	1	0	0	(.48, .71)	(.44, .74)	(.40, .78)
15	10	.67	67	51	34	19	8	2	0	0	(.55, .77)	(.51, .80)	(.47, .83)
15	11	.73	83	71	55	37	20	8	2	0	(.62, .83)	(.58, .86)	(.54, .88)
15	12	.80	93	87	75	60	40	21	7	1	(.69, .88)	(.65, .91)	(.61, .93)
15	13	.87	98	95	90	80	65	44	21	4	(.77, .93)	(.73, .95)	(.69, .97)
15	14	.93	100	99	97	94	86	72	49	19	(.85, .98)	(.82, .99)	(.77, .99)
15	15	1.00	100	100	100	99	97	93	81	56	(.93, 1.00)	(.90, 1.00)	(.87, 1.00)
16	8	.50	20	10	4	1	0	0	0	0	(.39, .61)	(.35, .65)	(.31, .69)
16	9	.56	36	21	10	4	1	0	0	0	(.45, .67)	(.41, .71)	(.37, .74)
16	10	.63	55	38	22	11	4	1	0	0	(.51, .73)	(.47, .76)	(.43, .80)
16	11	.69	74	58	40	23	11	3	0	0	(.57, .79)	(.54, .82)	(.49, .85)
16	12	.75	87	77	61	43	24	10	2	0	(.64, .84)	(.60, .87)	(.56, .89)
16	13	.81	95	90	80	65	45	24	8	1	(.71, .89)	(.67, .91)	(.63, .93)
16	14	.88	99	97	92	84	69	48	24	5	(.78, .94)	(.75, .95)	(.71, .97)
16	15	.94	100	99	98	95	88	75	52	21	(.86, .98)	(.83, .99)	(.79, .99)
16	16	1.00	100	100	100	99	98	94	83	58	(.94, 1.00)	(.91, 1.00)	(.87, 1.00)
17	9	.53	26	14	6	2	0	0	0	0	(.42, .64)	(.38, .67)	(.34, .71)
17	10	.59	44	27	14	6	2	0	0	0	(.48, .69)	(.44, .73)	(.40, .76)
17	11	.65	63	45	28	14	5	1	0	0	(.54, .75)	(.50, .78)	(.46, .81)
17	12	.71	79	65	47	28	13	4	1	0	(.60, .80)	(.56, .83)	(.52, .86)
17	13	.76	91	81	67	48	28	12	3	0	(.66, .86)	(.62, .87)	(.58, .90)
17	14	.82	97	92	84	69	50	28	10	1	(.73, .90)	(.69, .92)	(.65, .94)
17	15	.88	99	98	94	86	73	52	27	6	(.79, .94)	(.76, .96)	(.72, .97)
17	16	.94	100	100	99	96	90	78	55	23	(.87, .98)	(.84, .99)	(.80, .99)
17	17	1.00	100	100	100	99	98	95	85	60	(.94, 1.00)	(.91, 1.00)	(.88, 1.00)

Table A.2 (Continued)

Inferences about Universe Score Given
n and x for an Examinee

			Probability that π is at or above								Credibility Intervals		
n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
18	9	.50	19	9	3	1	0	0	0	0	(.39, .61)	(.36, .64)	(.32, .68)
18	10	.56	33	19	8	3	1	0	0	0	(.43, .66)	(.41, .69)	(.37, .73)
18	11	.61	51	33	18	8	2	0	0	0	(.50, .71)	(.47, .74)	(.43, .78)
18	12	.67	69	52	33	17	7	2	0	0	(.56, .76)	(.52, .79)	(.48, .82)
18	13	.72	84	70	53	33	16	5	1	0	(.62, .81)	(.58, .84)	(.54, .86)
18	14	.78	93	85	72	53	33	14	4	0	(.68, .86)	(.64, .88)	(.60, .90)
18	15	.83	98	94	87	74	54	32	11	1	(.74, .90)	(.71, .92)	(.67, .94)
18	16	.89	99	98	95	89	76	56	29	7	(.81, .95)	(.77, .96)	(.73, .98)
18	17	.94	100	100	99	97	92	80	58	25	(.87, .98)	(.85, .98)	(.81, .99)
18	18	1.00	100	100	100	100	99	95	86	62	(.94, 1.00)	(.92, 1.00)	(.89, 1.00)
19	10	.53	24	12	5	1	0	0	0	0	(.42, .63)	(.39, .66)	(.35, .70)
19	11	.58	40	24	11	4	1	0	0	0	(.47, .68)	(.44, .71)	(.40, .75)
19	12	.63	58	40	23	10	3	1	0	0	(.53, .73)	(.49, .76)	(.45, .79)
19	13	.68	75	58	39	21	9	2	0	0	(.58, .78)	(.55, .80)	(.50, .83)
19	14	.74	87	75	58	38	20	7	1	0	(.73, .82)	(.60, .85)	(.56, .87)
19	15	.79	95	88	76	59	37	17	4	0	(.69, .87)	(.66, .88)	(.62, .91)
19	16	.84	98	96	89	77	59	35	13	2	(.75, .91)	(.72, .93)	(.68, .94)
19	17	.89	100	99	96	91	79	60	32	8	(.81, .95)	(.78, .96)	(.75, .97)
19	18	.95	100	100	99	98	93	82	61	26	(.88, .98)	(.85, .99)	(.82, .99)
19	19	1.00	100	100	100	100	99	96	88	64	(.95, 1.00)	(.92, 1.00)	(.89, 1.00)
20	10	.50	17	8	3	1	0	0	0	0	(.40, .60)	(.36, .64)	(.33, .67)
20	11	.55	31	16	7	2	0	0	0	0	(.45, .65)	(.41, .68)	(.34, .72)
20	12	.60	48	29	15	6	1	0	0	0	(.50, .70)	(.46, .73)	(.42, .76)
20	13	.65	65	46	28	13	4	1	0	0	(.55, .74)	(.51, .77)	(.47, .80)
20	14	.70	80	64	45	26	11	3	0	0	(.60, .79)	(.57, .81)	(.53, .84)
20	15	.75	90	80	64	43	23	8	1	0	(.65, .83)	(.62, .86)	(.58, .88)
20	16	.80	96	91	80	63	41	20	5	0	(.71, .87)	(.67, .89)	(.64, .91)
20	17	.85	99	97	91	81	63	39	15	2	(.76, .91)	(.73, .93)	(.69, .94)
20	18	.90	100	99	97	93	82	63	35	8	(.82, .95)	(.79, .96)	(.76, .98)
20	19	.95	100	100	99	98	94	84	64	28	(.89, .98)	(.86, .99)	(.82, .99)
20	20	1.00	100	100	100	100	99	97	89	66	(.95, 1.00)	(.93, 1.00)	(.90, 1.00)

Table A.2 (Continued)

Inferences about Universe Score Given
n and x for an Examinee

			Probability that π is at or above								Credibility Intervals		
n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
21	11	0.52	23	11	4	1	0	0	0	0	(.42, .62)	(.39, .66)	(.35, .69)
21	12	0.57	38	21	9	3	1	0	0	0	(.47, .67)	(.44, .70)	(.40, .73)
21	13	0.62	55	35	19	7	2	0	0	0	(.52, .71)	(.48, .74)	(.45, .77)
21	14	0.67	71	53	33	16	6	1	0	0	(.57, .76)	(.53, .78)	(.49, .81)
21	15	0.71	84	70	51	30	13	4	0	0	(.62, .80)	(.58, .82)	(.54, .85)
21	16	0.76	93	84	69	48	27	10	2	0	(.67, .84)	(.64, .86)	(.60, .89)
21	17	0.81	97	93	84	68	46	23	6	0	(.72, .88)	(.69, .90)	(.62, .92)
21	18	0.86	99	98	93	84	67	42	17	2	(.77, .92)	(.74, .93)	(.71, .95)
21	19	0.90	100	99	98	94	85	66	38	9	(.83, .95)	(.80, .97)	(.77, .98)
21	20	0.95	100	100	100	99	95	86	66	30	(.89, .99)	(.87, .99)	(.83, 1.00)
21	21	1.00	100	100	100	100	99	97	90	68	(.95, 1.00)	(.93, 1.00)	(.90, 1.00)
22	11	0.50	16	7	2	0	0	0	0	0	(.40, .60)	(.37, .63)	(.34, .46)
22	12	0.55	29	14	5	1	0	0	0	0	(.45, .64)	(.41, .67)	(.38, .71)
22	13	0.59	44	26	12	4	1	0	0	0	(.49, .69)	(.46, .71)	(.42, .75)
22	14	0.64	61	41	23	10	3	0	0	0	(.54, .73)	(.50, .76)	(.47, .79)
22	15	0.68	76	59	38	20	7	2	0	0	(.58, .77)	(.55, .79)	(.51, .82)
22	16	0.73	88	75	56	35	16	5	1	0	(.63, .81)	(.60, .83)	(.56, .36)
22	17	0.77	95	87	73	53	31	12	2	0	(.68, .85)	(.65, .87)	(.61, .89)
22	18	0.82	98	94	86	72	50	26	7	0	(.73, .89)	(.70, .91)	(.66, .92)
22	19	0.86	99	98	95	86	70	46	19	3	(.78, .92)	(.76, .94)	(.72, .95)
22	20	0.91	100	100	98	95	87	69	41	11	(.84, .96)	(.81, .97)	(.78, .98)
22	21	0.95	100	100	100	99	96	88	68	32	(.90, .99)	(.87, .99)	(.84, 1.00)
22	22	1.00	100	100	100	100	99	98	91	69	(.95, 1.00)	(.93, 1.00)	(.90, 1.00)
23	12	0.52	21	9	3	1	0	0	0	0	(.42, .62)	(.39, .65)	(.36, .68)
23	13	0.57	35	18	7	2	0	0	0	0	(.47, .66)	(.44, .69)	(.40, .72)
23	14	0.61	51	31	15	5	1	0	0	0	(.51, .70)	(.48, .73)	(.44, .76)
23	15	0.65	67	47	27	12	4	1	0	0	(.56, .74)	(.52, .77)	(.49, .80)
23	16	0.70	81	64	44	23	9	2	0	0	(.60, .78)	(.57, .80)	(.53, .83)
23	17	0.74	90	79	61	39	19	6	1	0	(.65, .82)	(.62, .84)	(.58, .87)
23	18	0.78	96	90	77	58	34	14	3	0	(.69, .86)	(.66, .88)	(.63, .90)
23	19	0.83	99	96	89	75	54	29	9	1	(.74, .89)	(.71, .91)	(.68, .93)
23	20	0.87	100	99	96	88	74	50	21	3	(.79, .93)	(.76, .94)	(.73, .95)
23	21	0.91	100	100	99	96	89	72	44	12	(.84, .96)	(.82, .97)	(.79, .98)
23	22	0.96	100	100	100	99	97	89	71	34	(.90, .99)	(.88, .99)	(.85, 1.00)
23	23	1.00	100	100	100	100	100	98	92	71	(.96, 1.00)	(.94, 1.00)	(.91, 1.00)

AD-A116 424   AMERICAN COLL TESTING PROGRAM IOWA CITY IA   F/G 14/2
SOME STATISTICAL PROCEDURES FOR DOMAIN-REFERENCED TESTING: A HA--ETC(U)
FEB 81   R L BRENNAN   N00123-78-C-1206
UNCLASSIFIED   NPRDC-TN-81-6   NL

2 of 2
AD A
116424

END
DATE
FILMED
8-82
DTIC

Table A.2 (Continued)

n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
24	12	0.50	15	6	2	0	0	0	0	0	(.40, .60)	(.38, .62)	(.34, .66)
24	13	0.54	27	13	4	1	0	0	0	0	(.45, .64)	(.42, .66)	(.38, .70)
24	14	0.58	41	23	10	3	1	0	0	0	(.49, .68)	(.46, .70)	(.42, .73)
24	15	0.63	58	37	19	7	2	0	0	0	(.53, .71)	(.50, .74)	(.46, .77)
24	16	0.67	73	53	32	15	5	1	0	0	(.57, .75)	(.54, .78)	(.51, .81)
24	17	0.71	85	69	49	27	11	3	0	0	(.62, .79)	(.59, .81)	(.55, .84)
24	18	0.75	93	83	66	44	22	7	1	0	(.66, .83)	(.63, .85)	(.59, .87)
24	19	0.79	97	92	81	62	38	16	3	0	(.71, .86)	(.68, .88)	(.64, .90)
24	20	0.83	99	97	91	79	58	32	10	1	(.75, .90)	(.72, .91)	(.69, .93)
24	21	0.88	100	99	97	90	77	53	24	3	(.80, .93)	(.77, .94)	(.74, .96)
24	22	0.92	100	100	99	97	90	75	46	13	(.85, .96)	(.83, .97)	(.74, .98)
24	23	0.96	100	100	100	99	97	91	73	36	(.90, .99)	(.88, .99)	(.85, 1.00)
24	24	1.00	100	100	100	100	100	98	93	72	(.96, 1.00)	(.94, 1.00)	(.91, 1.00)
25	13	0.52	20	8	3	1	0	0	0	0	(.43, .61)	(.40, .64)	(.36, .67)
25	14	0.56	33	16	6	2	0	0	0	0	(.47, .65)	(.44, .68)	(.40, .71)
25	15	0.60	48	28	13	4	1	0	0	0	(.51, .69)	(.48, .72)	(.44, .75)
25	16	0.64	64	43	23	9	2	0	0	0	(.55, .73)	(.52, .75)	(.48, .78)
25	17	0.68	77	59	37	18	6	1	0	0	(.59, .76)	(.56, .79)	(.52, .81)
25	18	0.72	88	74	54	31	13	3	0	0	(.63, .80)	(.60, .82)	(.56, .85)
25	19	0.76	94	86	70	48	25	8	1	0	(.67, .83)	(.64, .85)	(.61, .88)
25	20	0.80	98	94	84	66	42	18	4	0	(.72, .87)	(.69, .89)	(.65, .91)
25	21	0.84	99	98	93	82	62	35	11	1	(.76, .90)	(.73, .92)	(.70, .93)
25	22	0.88	100	99	97	92	79	56	26	4	(.81, .93)	(.78, .95)	(.75, .96)
25	23	0.92	100	100	99	97	92	77	49	14	(.86, .96)	(.83, .97)	(.80, .98)
25	24	0.96	100	100	100	99	98	92	75	38	(.91, .99)	(.89, .99)	(.86, 1.00)
25	25	1.00	100	100	100	100	100	99	94	74	(.96, 1.00)	(.94, 1.00)	(.92, 1.00)
26	13	0.50	14	5	1	0	0	0	0	0	(.41, .59)	(.38, .62)	(.35, .65)
26	14	0.54	25	11	4	1	0	0	0	0	(.45, .63)	(.42, .66)	(.38, .69)
26	15	0.58	39	20	8	2	0	0	0	0	(.48, .67)	(.45, .69)	(.42, .72)
26	16	0.62	54	33	16	5	1	0	0	0	(.52, .70)	(.49, .73)	(.46, .76)
26	17	0.65	69	48	27	11	3	0	0	0	(.56, .74)	(.53, .76)	(.50, .79)
26	18	0.69	82	64	42	21	7	1	0	0	(.60, .77)	(.57, .80)	(.54, .82)
26	19	0.73	90	78	59	36	16	4	0	0	(.64, .81)	(.61, .83)	(.58, .85)
26	20	0.77	96	89	74	53	29	10	1	0	(.68, .84)	(.66, .86)	(.62, .88)
26	21	0.81	98	95	86	70	46	21	5	0	(.73, .87)	(.70, .89)	(.67, .91)
26	22	0.85	100	98	94	84	65	38	13	1	(.77, .91)	(.74, .92)	(.71, .94)
26	23	0.88	100	99	98	93	82	59	28	4	(.82, .94)	(.79, .95)	(.76, .96)
26	24	0.92	100	100	99	98	93	79	52	15	(.86, .96)	(.84, .97)	(.81, .98)
26	25	0.96	100	100	100	100	98	93	77	39	(.91, .99)	(.89, .99)	(.86, 1.00)
26	26	1.00	100	100	100	100	100	99	94	75	(.96, 1.00)	(.94, 1.00)	(.92, 1.00)

Table A.2 (Continued)

Inferences about Universe Score Given
n and x for an Examinee

			Probability that π is at or above								Credibility Intervals		
n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
27	14	0.52	19	7	2	0	0	0	0	0	(.43, .61)	(.40, .64)	(.37, .67)
27	15	0.56	30	14	5	1	0	0	0	0	(.46, .64)	(.44, .67)	(.40, .70)
27	16	0.59	45	25	10	3	0	0	0	0	(.50, .68)	(.47, .71)	(.44, .73)
27	17	0.63	60	38	19	7	1	0	0	0	(.54, .71)	(.51, .74)	(.48, .77)
27	18	0.67	74	54	32	14	4	1	0	0	(.58, .75)	(.55, .77)	(.51, .80)
27	19	0.70	85	69	47	25	9	2	0	0	(.62, .78)	(.59, .80)	(.55, .83)
27	20	0.74	93	82	64	40	18	5	0	0	(.66, .82)	(.63, .84)	(.59, .86)
27	21	0.78	97	91	78	57	32	12	2	0	(.70, .85)	(.67, .87)	(.63, .89)
27	22	0.81	99	96	89	74	50	24	6	0	(.74, .88)	(.71, .90)	(.68, .91)
27	23	0.85	100	99	95	86	69	41	14	1	(.78, .91)	(.75, .92)	(.72, .94)
27	24	0.89	100	100	98	94	84	62	31	5	(.82, .94)	(.80, .95)	(.77, .96)
27	25	0.93	100	100	100	98	94	81	54	16	(.87, .97)	(.84, .97)	(.81, .98)
27	26	0.96	100	100	100	100	98	94	78	41	(.91, .99)	(.89, .99)	(.87, 1.C0)
27	27	1.00	100	100	100	100	100	99	95	76	(.96, 1.00)	(.94, 1.00)	(.92, 1.00)
28	14	0.50	14	5	1	0	0	0	0	0	(.41, .59)	(.38, .62)	(.35, .65)
28	15	0.54	23	10	3	1	0	0	0	0	(.45, .62)	(.42, .65)	(.39, .68)
28	16	0.57	36	18	7	2	0	0	0	0	(.48, .66)	(.45, .68)	(.42, .71)
28	17	0.61	51	30	13	4	1	0	0	0	(.52, .69)	(.49, .72)	(.46, .75)
28	18	0.64	66	44	23	9	2	0	0	0	(.55, .73)	(.53, .75)	(.49, .78)
28	19	0.68	79	59	36	17	5	1	0	0	(.56, .76)	(.56, .78)	(.53, .81)
28	20	0.71	88	74	52	29	11	2	0	0	(.63, .79)	(.60, .81)	(.57, .84)
28	21	0.75	94	85	68	44	21	6	1	0	(.67, .82)	(.64, .81)	(.61, .86)
28	22	0.79	98	93	81	61	36	13	2	0	(.71, .85)	(.68, .87)	(.65, .89)
28	23	0.82	99	97	91	77	54	26	6	0	(.75, .88)	(.72, .90)	(.69, .92)
28	24	0.86	100	99	96	88	72	44	16	1	(.76, .91)	(.76, .93)	(.73, .94)
28	25	0.89	100	100	99	95	86	65	33	5	(.83, .94)	(.80, .95)	(.77, .96)
28	26	0.93	100	100	100	99	95	83	57	18	(.87, .97)	(.85, .98)	(.82, .98)
28	27	0.96	100	100	100	100	99	95	80	43	(.92, .99)	(.90, .99)	(.87, 1.00)
28	28	1.00	100	100	100	100	100	99	95	77	(.96, 1.00)	(.95, 1.00)	(.92, 1.00)

Table A.2 (Continued)

Inferences about Universe Score Given
n and x for an Examinee

			Probability that π is at or above								Credibility Intervals		
n	x	$\bar{x}$	.60	.65	.70	.75	.80	.85	.90	.95	67 Percent	80 Percent	90 Percent
29	15	0.52	18	7	2	0	0	0	0	0	(.43, .60)	(.40, .63)	(.37, .66)
29	16	0.55	29	13	4	1	0	0	0	0	(.46, .64)	(.44, .66)	(.40, .69)
29	17	0.59	42	22	8	2	0	0	0	0	(.50, .67)	(.47, .70)	(.44, .72)
29	18	0.62	57	35	16	5	1	0	0	0	(.53, .70)	(.51, .73)	(.47, .76)
29	19	0.66	71	49	27	11	3	0	0	0	(.57, .74)	(.54, .76)	(.51, .78)
29	20	0.69	82	64	41	20	6	1	0	0	(.60, .77)	(.58, .79)	(.54, .81)
29	21	0.72	91	78	57	33	13	3	0	0	(.64, .80)	(.61, .82)	(.58, .84)
29	22	0.76	96	88	72	49	24	7	1	0	(.68, .83)	(.65, .85)	(.62, .87)
29	23	0.79	98	94	84	65	39	15	3	0	(.72, .86)	(.69, .88)	(.66, .90)
29	24	0.83	99	98	92	80	57	29	7	0	(.75, .89)	(.73, .90)	(.70, .92)
29	25	0.86	100	99	97	90	74	48	18	2	(.79, .92)	(.77, .93)	(.74, .94)
29	26	0.90	100	100	99	96	88	68	35	6	(.83, .94)	(.81, .95)	(.78, .96)
29	27	0.93	100	100	100	99	96	85	59	19	(.88, .97)	(.85, .98)	(.83, .98)
29	28	0.97	100	100	100	100	99	95	82	45	(.92, .94)	(.90, .99)	(.87, 1.00)
29	29	1.00	100	100	100	100	100	99	96	79	(.96, 1.00)	(.95, 1.00)	(.93, 1.00)
30	15	0.50	13	4	1	0	0	0	0	0	(.41, .59)	(.39, .61)	(.36, .64)
30	16	0.53	22	9	2	0	0	0	0	0	(.45, .62)	(.42, .64)	(.39, .67)
30	17	0.57	34	16	5	1	0	0	0	0	(.48, .65)	(.45, .68)	(.42, .71)
30	18	0.60	48	26	11	3	0	0	0	0	(.51, .68)	(.49, .71)	(.45, .74)
30	19	0.63	62	40	19	6	1	0	0	0	(.55, .71)	(.52, .74)	(.49, .76)
30	20	0.67	75	54	31	13	3	0	0	0	(.58, .74)	(.55, .77)	(.52, .79)
30	21	0.70	86	69	46	23	7	1	0	0	(.62, .78)	(.59, .80)	(.56, .82)
30	22	0.73	93	81	61	37	15	3	0	0	(.65, .81)	(.63, .83)	(.59, .85)
30	23	0.77	97	90	76	53	27	8	1	0	(.69, .83)	(.66, .85)	(.63, .87)
30	24	0.80	99	95	87	69	43	17	3	0	(.72, .86)	(.70, .88)	(.67, .90)
30	25	0.83	100	98	94	82	61	32	8	0	(.76, .89)	(.74, .91)	(.71, .92)
30	26	0.87	100	99	98	92	77	51	19	2	(.80, .92)	(.78, .93)	(.75, .95)
30	27	0.90	100	100	99	97	89	70	38	7	(.84, .94)	(.82, .96)	(.79, .97)
30	28	0.93	100	100	100	99	96	86	61	20	(.88, .97)	(.86, .98)	(.83, .98)
30	29	0.97	100	100	100	100	99	96	83	46	(.92, .99)	(.90, .99)	(.88, 1.00)
30	30	1.00	100	100	100	100	100	99	96	80	(.97, 1.00)	(.95, 1.00)	(.93, 1.00)

Table A.3

Probability that Two Standard Normal
Variables, with Correlation Equal to KR-21,
are Both Less Than or Equal to z

	KR-21																*****	
z	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00	z
-1.95	0.002	0.002	0.002	0.003	0.003	0.004	0.005	0.006	0.006	0.007	0.009	0.010	0.011	0.013	0.015	0.018	0.026	-1.95
-1.90	0.002	0.002	0.003	0.004	0.004	0.005	0.006	0.006	0.007	0.009	0.010	0.011	0.013	0.015	0.017	0.021	0.029	-1.90
-1.85	0.002	0.003	0.004	0.004	0.005	0.006	0.007	0.008	0.009	0.010	0.011	0.013	0.015	0.017	0.020	0.023	0.032	-1.85
-1.80	0.003	0.004	0.004	0.005	0.006	0.007	0.008	0.009	0.010	0.011	0.013	0.015	0.017	0.019	0.022	0.026	0.036	-1.80
-1.75	0.004	0.004	0.005	0.006	0.007	0.008	0.009	0.010	0.012	0.013	0.015	0.017	0.019	0.022	0.025	0.029	0.040	-1.75
-1.70	0.004	0.005	0.006	0.007	0.008	0.009	0.010	0.012	0.013	0.015	0.017	0.019	0.022	0.025	0.028	0.033	0.045	-1.70
-1.65	0.005	0.006	0.007	0.008	0.009	0.011	0.012	0.014	0.015	0.017	0.019	0.022	0.024	0.028	0.031	0.037	0.049	-1.65
-1.60	0.006	0.007	0.008	0.009	0.011	0.012	0.014	0.016	0.018	0.020	0.022	0.025	0.028	0.031	0.035	0.041	0.055	-1.60
-1.55	0.007	0.008	0.010	0.011	0.013	0.014	0.016	0.018	0.020	0.022	0.025	0.028	0.031	0.035	0.039	0.046	0.061	-1.55
-1.50	0.009	0.010	0.011	0.013	0.015	0.016	0.018	0.020	0.023	0.025	0.028	0.031	0.035	0.039	0.044	0.051	0.067	-1.50
-1.45	0.010	0.012	0.013	0.015	0.017	0.019	0.021	0.023	0.026	0.029	0.032	0.035	0.039	0.044	0.049	0.056	0.074	-1.45
-1.40	0.012	0.014	0.015	0.017	0.019	0.022	0.024	0.027	0.029	0.032	0.036	0.039	0.044	0.048	0.054	0.062	0.081	-1.40
-1.35	0.014	0.016	0.018	0.020	0.022	0.025	0.027	0.030	0.033	0.036	0.040	0.044	0.049	0.054	0.060	0.068	0.089	-1.35
-1.30	0.016	0.018	0.021	0.023	0.025	0.028	0.031	0.034	0.037	0.041	0.045	0.049	0.054	0.060	0.066	0.075	0.097	-1.30
-1.25	0.019	0.021	0.024	0.026	0.029	0.032	0.035	0.038	0.042	0.046	0.050	0.055	0.060	0.066	0.073	0.083	0.106	-1.25
-1.20	0.022	0.024	0.027	0.030	0.033	0.036	0.040	0.043	0.047	0.051	0.056	0.061	0.066	0.073	0.081	0.091	0.115	-1.20
-1.15	0.025	0.028	0.031	0.034	0.037	0.041	0.045	0.048	0.053	0.057	0.062	0.067	0.073	0.080	0.088	0.099	0.125	-1.15
-1.10	0.029	0.032	0.035	0.039	0.042	0.046	0.050	0.054	0.059	0.064	0.069	0.075	0.081	0.088	0.097	0.108	0.136	-1.10
-1.05	0.033	0.037	0.040	0.044	0.048	0.052	0.056	0.061	0.065	0.071	0.076	0.082	0.089	0.097	0.106	0.118	0.147	-1.05
-1.00	0.038	0.042	0.045	0.049	0.054	0.058	0.063	0.067	0.073	0.078	0.084	0.090	0.098	0.106	0.115	0.128	0.159	-1.00
-0.95	0.043	0.047	0.051	0.056	0.060	0.065	0.070	0.075	0.080	0.086	0.092	0.099	0.107	0.115	0.126	0.139	0.171	-0.95
-0.90	0.049	0.053	0.058	0.062	0.067	0.072	0.077	0.083	0.089	0.095	0.102	0.109	0.117	0.126	0.137	0.150	0.184	-0.90
-0.85	0.056	0.060	0.065	0.070	0.075	0.080	0.086	0.092	0.098	0.104	0.111	0.119	0.127	0.137	0.148	0.163	0.198	-0.85
-0.80	0.063	0.068	0.073	0.078	0.083	0.089	0.095	0.101	0.107	0.114	0.122	0.130	0.138	0.148	0.160	0.175	0.212	-0.80
-0.75	0.071	0.076	0.081	0.087	0.092	0.098	0.104	0.111	0.118	0.125	0.133	0.141	0.150	0.160	0.173	0.189	0.227	-0.75
-0.70	0.079	0.085	0.090	0.096	0.102	0.108	0.115	0.122	0.129	0.136	0.144	0.153	0.163	0.173	0.186	0.202	0.242	-0.70
-0.65	0.088	0.094	0.100	0.106	0.112	0.119	0.126	0.133	0.140	0.148	0.157	0.166	0.176	0.187	0.200	0.217	0.258	-0.65
-0.60	0.098	0.104	0.111	0.117	0.124	0.130	0.138	0.145	0.153	0.161	0.170	0.179	0.189	0.201	0.214	0.232	0.274	-0.60
-0.55	0.109	0.115	0.122	0.129	0.136	0.143	0.150	0.158	0.166	0.174	0.183	0.193	0.204	0.216	0.230	0.248	0.291	-0.55
-0.50	0.121	0.127	0.134	0.141	0.148	0.156	0.163	0.171	0.180	0.188	0.198	0.208	0.219	0.231	0.245	0.264	0.309	-0.50

Table A.3 (Continued)

Probability that Two Standard Normal
Variables, with Correlation Equal to KR-21,
are Both Less Than or Equal to z

	KR-21																
z	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	***** 1.00	z
-0.45	0.133	0.140	0.147	0.154	0.162	0.169	0.177	0.185	0.194	0.203	0.213	0.223	0.234	0.247	0.262	0.281	0.326	-0.45
-0.40	0.146	0.154	0.161	0.168	0.176	0.184	0.192	0.200	0.209	0.218	0.228	0.239	0.250	0.263	0.278	0.298	0.345	-0.40
-0.35	0.161	0.168	0.175	0.183	0.191	0.199	0.207	0.216	0.225	0.234	0.245	0.255	0.267	0.280	0.296	0.316	0.363	-0.35
-0.30	0.176	0.183	0.191	0.199	0.207	0.215	0.223	0.232	0.241	0.251	0.261	0.272	0.284	0.298	0.314	0.334	0.382	-0.30
-0.25	0.191	0.199	0.207	0.215	0.223	0.232	0.240	0.249	0.259	0.268	0.279	0.290	0.302	0.316	0.332	0.352	0.401	-0.25
-0.20	0.208	0.216	0.224	0.232	0.240	0.249	0.258	0.267	0.276	0.286	0.297	0.308	0.320	0.334	0.350	0.371	0.421	-0.20
-0.15	0.225	0.233	0.241	0.250	0.258	0.267	0.276	0.285	0.295	0.305	0.315	0.327	0.339	0.353	0.369	0.390	0.440	-0.15
-0.10	0.244	0.252	0.260	0.268	0.277	0.285	0.294	0.304	0.313	0.324	0.334	0.346	0.358	0.372	0.389	0.410	0.460	-0.10
-0.05	0.262	0.271	0.279	0.287	0.296	0.305	0.314	0.323	0.333	0.343	0.354	0.365	0.378	0.392	0.408	0.430	0.480	-0.05
0.0	0.282	0.290	0.298	0.307	0.315	0.324	0.333	0.343	0.352	0.363	0.373	0.385	0.398	0.412	0.428	0.449	0.500	0.0
0.05	0.302	0.310	0.319	0.327	0.336	0.344	0.354	0.363	0.373	0.383	0.394	0.405	0.418	0.432	0.448	0.469	0.520	0.05
0.10	0.323	0.331	0.339	0.348	0.356	0.365	0.374	0.383	0.393	0.403	0.414	0.425	0.438	0.452	0.468	0.490	0.540	0.10
0.15	0.345	0.353	0.361	0.369	0.377	0.386	0.395	0.404	0.414	0.424	0.435	0.446	0.458	0.472	0.489	0.510	0.560	0.15
0.20	0.366	0.374	0.382	0.391	0.399	0.407	0.416	0.425	0.435	0.445	0.455	0.467	0.479	0.493	0.509	0.530	0.579	0.20
0.25	0.389	0.396	0.404	0.412	0.421	0.429	0.438	0.447	0.456	0.466	0.476	0.487	0.500	0.513	0.529	0.550	0.599	0.25
0.30	0.411	0.419	0.427	0.435	0.443	0.451	0.459	0.468	0.477	0.487	0.497	0.508	0.520	0.534	0.549	0.570	0.618	0.30
0.35	0.434	0.442	0.449	0.457	0.465	0.473	0.481	0.490	0.499	0.508	0.518	0.529	0.541	0.554	0.569	0.589	0.637	0.35
0.40	0.457	0.464	0.472	0.479	0.487	0.495	0.503	0.511	0.520	0.529	0.539	0.550	0.561	0.574	0.589	0.609	0.655	0.40
0.45	0.480	0.487	0.494	0.502	0.509	0.517	0.525	0.533	0.541	0.550	0.560	0.570	0.581	0.594	0.609	0.628	0.674	0.45

Table A.3 (Continued)

Probability that Two Standard Normal
Variables, with Correlation Equal to KR-21,
are Both Less Than or Equal to z

	KR-21																	
																	*****	
z	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00	z
0.50	0.504	0.510	0.517	0.524	0.531	0.539	0.546	0.554	0.562	0.571	0.581	0.591	0.601	0.614	0.628	0.647	0.691	0.50
0.55	0.527	0.533	0.540	0.546	0.553	0.560	0.568	0.575	0.583	0.592	0.601	0.611	0.621	0.633	0.647	0.665	0.709	0.55
0.60	0.550	0.556	0.562	0.569	0.575	0.582	0.589	0.596	0.604	0.612	0.621	0.630	0.641	0.652	0.666	0.684	0.726	0.60
0.65	0.573	0.578	0.584	0.590	0.597	0.603	0.610	0.617	0.625	0.632	0.641	0.650	0.660	0.671	0.684	0.701	0.742	0.65
0.70	0.595	0.601	0.606	0.612	0.618	0.624	0.631	0.638	0.645	0.652	0.660	0.669	0.679	0.689	0.702	0.719	0.758	0.70
0.75	0.617	0.622	0.628	0.633	0.639	0.645	0.651	0.658	0.664	0.672	0.679	0.688	0.697	0.707	0.719	0.735	0.773	0.75
0.80	0.639	0.644	0.649	0.654	0.660	0.665	0.671	0.677	0.684	0.690	0.698	0.706	0.715	0.725	0.736	0.752	0.788	0.80
0.85	0.660	0.665	0.670	0.674	0.680	0.685	0.690	0.696	0.702	0.709	0.716	0.723	0.732	0.741	0.753	0.767	0.802	0.85
0.90	0.681	0.685	0.690	0.694	0.699	0.704	0.709	0.715	0.721	0.727	0.733	0.741	0.749	0.758	0.768	0.782	0.816	0.90
0.95	0.701	0.705	0.709	0.713	0.718	0.723	0.728	0.733	0.738	0.744	0.750	0.757	0.765	0.773	0.784	0.797	0.829	0.95
1.00	0.721	0.724	0.728	0.732	0.736	0.741	0.745	0.750	0.755	0.761	0.767	0.773	0.780	0.788	0.798	0.811	0.841	1.00
1.05	0.740	0.743	0.746	0.750	0.754	0.758	0.762	0.767	0.772	0.777	0.782	0.788	0.795	0.803	0.812	0.824	0.853	1.05
1.10	0.758	0.761	0.764	0.767	0.771	0.775	0.779	0.783	0.787	0.792	0.797	0.803	0.810	0.817	0.826	0.837	0.864	1.10
1.15	0.775	0.778	0.781	0.784	0.787	0.791	0.794	0.798	0.803	0.807	0.812	0.817	0.823	0.830	0.838	0.849	0.875	1.15
1.20	0.792	0.794	0.797	0.800	0.803	0.806	0.809	0.813	0.817	0.821	0.826	0.831	0.836	0.843	0.850	0.860	0.885	1.20
1.25	0.808	0.810	0.812	0.815	0.818	0.821	0.824	0.827	0.831	0.835	0.839	0.844	0.849	0.855	0.862	0.871	0.894	1.25
1.30	0.823	0.825	0.827	0.829	0.832	0.834	0.837	0.840	0.844	0.847	0.851	0.856	0.861	0.866	0.873	0.882	0.903	1.30
1.35	0.837	0.839	0.841	0.843	0.845	0.848	0.850	0.853	0.856	0.859	0.863	0.867	0.872	0.877	0.883	0.891	0.911	1.35
1.40	0.850	0.852	0.854	0.856	0.858	0.860	0.862	0.865	0.868	0.871	0.874	0.878	0.882	0.887	0.893	0.900	0.919	1.40
1.45	0.863	0.865	0.866	0.868	0.870	0.872	0.874	0.876	0.879	0.882	0.885	0.888	0.892	0.896	0.902	0.909	0.926	1.45
1.50	0.875	0.876	0.878	0.879	0.881	0.883	0.885	0.887	0.889	0.892	0.895	0.898	0.901	0.905	0.910	0.917	0.933	1.50
1.55	0.886	0.887	0.889	0.890	0.891	0.893	0.895	0.897	0.899	0.901	0.904	0.907	0.910	0.914	0.918	0.924	0.939	1.55
1.60	0.897	0.898	0.899	0.900	0.901	0.903	0.904	0.906	0.908	0.910	0.912	0.915	0.918	0.921	0.926	0.931	0.945	1.60
1.65	0.906	0.907	0.908	0.909	0.910	0.912	0.913	0.915	0.916	0.918	0.920	0.923	0.926	0.929	0.933	0.938	0.951	1.65
1.70	0.915	0.916	0.917	0.918	0.919	0.920	0.921	0.923	0.924	0.926	0.928	0.930	0.932	0.935	0.939	0.944	0.955	1.70
1.75	0.923	0.924	0.925	0.926	0.927	0.928	0.929	0.930	0.931	0.933	0.935	0.937	0.939	0.942	0.945	0.949	0.960	1.75
1.80	0.931	0.932	0.932	0.933	0.934	0.935	0.936	0.937	0.938	0.940	0.941	0.943	0.945	0.947	0.950	0.954	0.964	1.80
1.85	0.938	0.939	0.939	0.940	0.941	0.941	0.942	0.943	0.944	0.946	0.947	0.949	0.950	0.953	0.955	0.959	0.968	1.85
1.90	0.945	0.945	0.946	0.946	0.947	0.947	0.948	0.949	0.950	0.951	0.952	0.954	0.955	0.957	0.960	0.963	0.971	1.90
1.95	0.950	0.951	0.951	0.952	0.952	0.953	0.954	0.954	0.955	0.956	0.957	0.959	0.960	0.962	0.964	0.967	0.974	1.95

Appendix B

Technical Notes

These notes are provided for two reasons: (a) to cite appropriate technical background and references for each section of the handbook; and (b) to provide a limited amount of technical justification for equations and/or procedures that are not specifically reported in readily available references. However, there is no intent to cite all potentially relevant references or to verify _in detail_ all equations and/or procedures.

In the body of this handbook, distinctions have been drawn only very rarely between parameters and estimates of parameters. In these technical notes such distinctions are made through the use of a "hat" (^) above unbiased estimates of parameters, which are denoted by Greek letters. The reader should be careful not to confuse this use of a "hat" with the use already made of this symbol in the body of the handbook. Specifically, the "hat" symbol is also used to distinguish between the _sample_ variances s^2 and $\hat{s}^2$, where the former involves a denominator of $\underline{n}$ and the latter involves a denominator of $\underline{n} - 1$. (Of course, $\hat{s}^2$ is an unbiased estimate of a parameter, but usually not a parameter of interest, here.)

Section 1

Berk (1980) provides an edited book of readings on the subject of domain-referenced (or criterion-referenced) measurements. Most of the topics treated in this handbook are also covered in Berk (1980). Also, Hambleton, Swaminathan, Algina, and Coulson (1978) provide a technical review of many issues treated here; Millman (1979) provides a brief review written principally for practitioners; and Nitko (1980) reviews the many varieties of criterion-referenced tests. It should be noted, however, that there are clear differences between this handbook and the above references--differences in emphasis and scope, as well as occasional differences in perspective and approach.

Many introductory measurement textbooks give considerable attention to defining objectives and tables of specifications. Recently, Ellis and Wulfeck (1979) and Ellis, Wulfeck, and Fredericks (1979) have developed a task/content matrix for specific use in Navy training that involves domain-referenced testing.

Section 2

Most introductory measurement textbooks provide detailed discussion of item analysis procedures. Even though such discussions usually emphasize norm-referenced testing, many of the guidelines typically suggested are relevant for domain-referenced testing, too--with one noticeable exception. In the opinion of this author, it is not generally a good practice in domain-referenced testing to select items in a systematic manner so as to obtain some pre-specified distribution of item difficulty levels and/or discrimination indices. More specifically, this is not a good practice if a test is to be used solely for the purpose of making domain-referenced interpretations of test scores.

The discrimination index, B, discussed in Section 2 is treated by Brennan (1972). More recently, Harris and Wilcox (1980) have commented on this index.

Section 3

The procedure suggested in Section 3 for establishing a cutting score is a slight modification of a procedure originally proposed by Angoff (1971); and the developments involving $\sigma(\bar{y})$ are discussed by Brennan and Lockwood (1980). The specific equations for $\sigma(\bar{y})$ in Table 3.2 can be derived in the manner outlined below.

Let the probability assigned by rater *r* (r=1, 2, ..., t) to item *i* (i=1, 2, ..., m) for a set of *m* items be:

$$y_{ri} = \lambda + \lambda_{r}\sim + \lambda_{i}\sim + \lambda_{ri}\sim$$

where λ is the grand mean and the $\lambda\sim$ are score effects as discussed by Brennan and Lockwood (1980). It can be shown that unbiased estimates of the variance of these score effects, in terms of the sample statistics reported in Table 3.2, are:

$$\hat{\sigma}^2(ri) = [\sum_{r} \hat{s}_i^2(y_{ri}) - t\ \hat{s}^2(\bar{y}_i)]/(t-1) \qquad \text{(B1)}$$

$$\hat{\sigma}^2(r) = \hat{s}^2(\bar{y}_r) - \hat{\sigma}^2(ri)/m \qquad \text{(B2)}$$

$$\hat{\sigma}^2(i) = \hat{s}^2(\bar{y}_i) - \hat{\sigma}^2(ri)/t\ . \qquad \text{(B3)}$$

For random samples of t raters and random samples of n items (n need not equal m) an unbiased estimate of $\sigma^2(\bar{y})$ is:

$$\hat{\sigma}^2(\bar{y}) = \frac{\hat{\sigma}^2(r)}{t} + \frac{\hat{\sigma}^2(i)}{n} + \frac{\hat{\sigma}^2(ri)}{nt} \quad . \tag{B4}$$

Using Equations B1 to B3 in B4 we obtain

$$\hat{\sigma}^2(\bar{y}) = \frac{\hat{s}^2(\bar{y}_i)}{n} + \frac{\hat{s}^2(\bar{y}_r)}{t} - \left[\frac{\sum_r \hat{s}_i^2(y_{ri})}{mt(t-1)} - \frac{\hat{s}^2(\bar{y}_i)}{m(t-1)}\right] , \tag{B5}$$

where the bracketed term in Equation B5 is $\hat{\sigma}^2(ri)/tm$, which constitutes the A-term defined in Table 3.2. The square root of Equation B5 is Equation 3.2 in Table 3.2; and when n equals m, the square root of Equation B5 is Equation 3.1 in Table 3.2.

Finally, as $n \to \infty$, it is evident from Equation B4 that

$$\hat{\sigma}^2(\bar{y}) = \hat{\sigma}^2(r)/t \quad ;$$

and using Equation B2,

$$\begin{aligned} \hat{\sigma}^2(\bar{y}) &= \hat{s}^2(\bar{y}_r)/t - \hat{\sigma}^2(ri)/mt \\ &= \hat{s}^2(\bar{y}_r)/t - A \quad . \end{aligned} \tag{B6}$$

The square root of $\hat{\sigma}^2(\bar{y})$ in Equation B6 is Equation 3.3 in Table 3.2.

Section 4

Table A.1, which is discussed in Section 4, results from applying a minimax procedure presented in Huynh (1980, pp. 170-171). As such this procedure is basically an extension of an approach suggested by Fhanér (1974) and treated by Wilcox (1976). It should be noted, however, that where Huynh talks about the loss ratio Q, this author talks about 1/Q; e.g., if false positive errors are twice as serious as false negative errors, Huynh says the loss ratio is Q = .50, and in Section 4 this loss ratio is identified as 1/.50 = 2. Of course, this difference is simply a question of definition.

It is suggested in Section 4 that a confidence interval for π_o from a cutting score study be considered as one possible way to define an indifference zone. In doing so, it might be argued that one is implicitly violating the assumption of 0 - 1 referral loss, which is an assumption made by Huynh (1980) in his formulation of the minimax procedure used to generate Table A.1. Another approach that might be considered is to eliminate the indifference zone and use, $\bar{y}$ and $\sigma(\bar{y})$ from a cutting score study to establish an ogive-shaped referral success function, but this is considerably more complicated than the approach taken in this handbook.

Section 5

With respect to technical issues, Section 5 is based principally on Table A.2 which was developed under the assumptions of a binomial likelihood and a uniform beta prior distribution for π (sometimes called a non-informative prior).

Specifically, an entry in the left-hand part of Table A.2 is:

$$\text{Prob}\ (\pi_p \geq \pi | n,\ x_p) \ = \ 1 - I_{\pi}(x_p + 1,\ n - x_p + 1)\ ;$$

where x_p and π_p are an examinee's observed and universe scores, respectively; and $I_{\pi}(x_p + 1, n - x_p + 1)$ is the incomplete beta function with parameters $x_p + 1$ and $n - x_p + 1$. An entry in the right-hand part of Table A.2 is a Bayesian credibility interval for π under the assumption of a uniform beta prior distribution. Technically, these intervals are called highest density regions. (Some might quarrel with calling an interval a highest density region when $n = x$.) Readers unfamiliar with these Bayesian concepts can consult Novick and Jackson (1974, Chapter 5.)

A principal reason for using a beta prior here is that this assumption results in a Bayesian credibility interval, which enables one to make probability statements about the parameter π. By contrast, a confidence interval allows one to make probability statements about intervals covering π. Some might argue that in specific contexts, a uniform beta prior is frequently unrealistic because a decision-maker may know a great deal about an examinee. However, to assess "informative" (i.e., non-uniform) beta priors in a decision-making process virtually necessitates an interactive computing system such as CADA (Isaacs and Novick, 1978). Furthermore, a decision-maker would need to justify the specific "informative" prior chosen in each and every individual case.

It should be noted that the Fhanér-Wilcox-Huynh approach to establishing an advancement score, discussed in Section 4, involves consideration of false positive and false negative errors, but a uniform beta prior is *not* assumed in their approach. There is, therefore, a degree of discontinuity between Sections 4 and 5. (For the purpose of establishing an advancement score, a *uniform* beta prior assumption for a *group* of examinees seems highly unrealistic to this author. One might argue that an informative beta prior could be used, but, as indicated previously, the process of doing so is far from trivial and clearly beyond the scope of this handbook.)

Section 6

The theoretical framework used in Section 6 for integrating squared error loss and threshold loss approaches is provided by Kane and Brennan (1980). In addition, a considerable number of papers have been published that involve consideration of one loss function or the other.

Concerning threshold loss, the following publications, among others, are relevant: (a) Hambleton and Novick (1973) provided the first integrated treatment of threshold loss and domain-referenced testing issues; (b) Swaminathan, Hambleton, and Algina (1974) suggested using coefficient Kappa; (c) Huynh (1976) and Subkoviak (1976) provided procedures for estimating threshold loss coefficients based on a single test; and (d) Subkoviak (1980) has reviewed much of the work in this area.

Concerning squared error loss, the following publications, among others, are relevant: (a) using classical test theory assumptions, Livingston (1972) proposed a reliability-like coefficient for domain-referenced tests; (b) using generalizability theory, Brennan and Kane (1977 a, b) proposed two coefficients and a definition of error variance; (c) Brennan (1979a) has provided a computer program for performing computations involving squared error loss considerations with domain-referenced testing; and (d) Brennan (1980b) has reviewed much of the work in this area.

The formulas in Table 6.3 are computationally easy to use, but they are rather unusual expressions for estimates of their respective paramaters. For this reason, the derivations of these expressions are briefly outlined below.

Let the observed score for person p (p=1, 1, ..., k) on item i (i=1, 2, ..., n) be:

$$X_{pi} = \mu + \pi_{p}\sim + \beta_{i}\sim + \pi\beta_{pi} \;;$$

where μ is the grand mean in the population of persons and universe of items; $\pi_{p}\sim$ is the score effect for person p ($\pi_{p} = \mu + \pi_{p}\sim$); $\beta_{i}\sim$ is the score effect for item i; and $\pi\beta_{pi}\sim$ is the effect for the interaction of person p and item i, which is confounded with experimental error. (See Brennan and Kane, 1977 a, for more detail.)

It is well-known that an unbiased estimate of $\sigma^2(\pi)$ is:

$$\hat{\sigma}^2(\pi) = [MS(p) - MS(pi)]/k \qquad (B7)$$

where "MS" is "mean square"; and, it is relatively easy to show that, for dichotomous data, Equation B6 can be expressed as Equation 6.1 in Table 6.3. In a similar manner, it can be shown that

$$\hat{\sigma}^2(\beta) = \frac{n[k\ s^2(\bar{x}_i) + s^2(\bar{x}_p) - \bar{x}(1-\bar{x})]}{(n-1)(k-1)} \tag{B8}$$

and

$$\hat{\sigma}^2(\pi\beta) = \frac{n\ k\ [\bar{x}(1-\bar{x}) - s^2(\bar{x}_p) - s^2(\bar{x}_i)]}{(n-1)(k-1)} \ . \tag{B9}$$

Now,

$$\hat{\sigma}^2(\Delta) = [\hat{\sigma}^2(\beta) + \hat{\sigma}^2(\pi\beta)]/n \quad ; \tag{B10}$$

and replacement of Equations B8 and B9 in B10 gives (after simplifying terms) Equation 6.2 in Table 6.3.

Brennan and Kane (1977a) report that a consistent estimate of $\Phi(c_o)$ is:

$$\hat{\Phi}(c_o) = 1 - \frac{1}{n-1}\left[\frac{\bar{x}(1-\bar{x}) - s^2(\bar{x}_p)}{(\bar{x}-c_o)^2 + s^2(\bar{x}_p)}\right] \tag{B11}$$

$$= 1 - \left\{\frac{[\bar{x}(1-\bar{x}) - s^2(\bar{x}_p)]/(n-1)}{(\bar{x}-c_o)^2 + s^2(\bar{x}_p)}\right\} \ . \tag{B12}$$

The numerator of the term in braces is simply $\sigma^2(\Delta)$ given by Equation 6.2 in Table 6.3; consequently, Equation B11 can be expressed as Equation 6.3

in Table 6.3. [Technically, $\sigma^2(\Delta)$ in Equation 6.3 should be $\hat{\sigma}^2(\Delta)$; but, as previously stated, notational distinctions between parameters and estimates are not made in the body of this handbook.] Equation 6.4 follows from the fact that $\hat{\Phi}(c_o)$ equals KR-21 if $c_o = \bar{x}$ (see Brennan, 1977). The expression for KR-21 in Equation 6.3 may appear strange because it involves $\hat{\sigma}^2(\Delta)$, but it is easily verified that this expression is algebraically identical to the well-known expression for KR-21.

The steps provided in Table 6.5 for obtaining estimates of threshold loss coefficients of agreement are based on Huynh's (1976) normal approxmimation procedure (see, also, Subkoviak, 1980), with<u>out</u> using an arcsine transformation (see Peng & Subkoviak, in press). In Table 6.5 reference is made to using the "closest" value in Table A.3; alternatively, one can obtain better estimates using linear interpolation (see Huynh, 1978--different context, but same process). Huynh (1978) provides a computer program for estimating threshold loss coefficients; as well as tables of estimates of p_o , Kappa, and their standard errors for test lengths of 5 to 10 items (see, also, Huynh & Saunders, 1980).

Since the procedure outlined in Table 6.4 is based on a normal approximation, estimates obtained using this procedure may be somewhat biased. However, the degree of bias is likely to be small unless n is quite small and/or c_o is quite close to one.

In Table 6.5, Equation 6.6 is simply $[\hat{\sigma}^2(\beta) + \hat{\sigma}^2(\pi\beta)]/n'$; and the remaining equations and steps constitute a somewhat ad hoc approach for using Huynh's normal approximation procedure to estimate the proportion of inconcsistent decisions for a test of length n'.

Brennan and Kane (1977b) show that $\hat{\sigma}^2(\Delta)$ is algebraically equal to the average of the squared values of $\hat{\sigma}(\Delta_p)$ in Table 5.4. Note also that $\hat{\sigma}(\Delta_p)$ is identical to Lord's (1957) formula for the standard error of measurement of an examinee's mean score.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L Thorndike (Ed.), Educational Measurement. Washington, DC: American Council on Education, 1971.

Berk, R. E. (Ed.). Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Brennan, R. L. A generalized uper-lower item discrimination index. Educational and Psychological Measurement, 1972, 32, 289-303.

Brennan, R. L. KR-21 and lower limits of an index of dependability for mastery tests (ACT Tech. Bulletin No. 27). Iowa City, IA: The American College Testing Program, December 1977.

Brennan, R. L. Handbook for GAPID: A FORTRAN IV computer program for generalizability analyses with single-facet designs (NPRDC Tech. Note 80-13). San Diego: Navy Personnel Research and Development Center, May 1980. (a)

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980. (b)

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika, 1977, 42, 609-625; Errata, 1978, 43, 289. (b)

Brennan, R. L., & Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 1980, 4, 219-240.

Ellis, J. A., & Wulfeck, W. H. Assessing objective-test consistency: A systematic procedure for constructing criterion-referenced tests. (NPRDC SR 80-15). San Diego: Navy Personnel Research and Development Center, 1979.

Ellis, J. A., Wulfeck, W. H., & Fredericks, P. S. The instructional quality inventory--II. User's manual (NPRDC SR 79-24). San Diego: Navy Personnel Research and Development Center, 1979.

Fhanér, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harris, C. W., & Wilcox, R. Brennan's B is Pierce's theta. Educational and Psychological Measurement, 1980, 40, 307-311.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. *Computation and inference for two reliability indices in mastery testing based on the beta-binomial model*. Publication Series in Mastery Testing, Research Memorandum 78-1. Columbia, SC: College of Education, University of South Carolina, 1978.

Huynh, H. A non-randomized minimax solution for passing scores in the binomial error model. *Psychometrika*, 1980, *45*, 167-182.

Huynh, H., & Saunders, J. C. *Solutions for some technical problems in domain-referenced mastery testing* (Final Report). Columbia, SC: Department of Educational Research and Psychology, College of Education, University of South Carolina, August 1980.

IMSL Libraries (7th Edition). Houston, International Mathematical and Statistical Libraries, Inc., July 1979.

Isaacs, G. L., Christ, D. E., Novick, M. R., & Jackson, P. H. *Tables for Bayesian statisticians*. Iowa City, IA: The University of Iowa, 1974.

Isaacs, G. L., & Novick, M. R. *Manual for the computer-assisted data analysis monitor--1978*. Iowa City, IA: Iowa Testing Programs, 1978.

Kane, M. T., & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 1980, *4*, 105-126.

Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26.

Lord, F. M. Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, 1957, *17*, 510-521.

Millman, J. Reliability and validity of criterion-referenced test scores. In R. E. Traub (Ed.), New directions for testing and measurement: Methodological developments (No. 4). San Francisco: Jossey-Bass, 1979.

Nitko, A. J. Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 1980, 50, 461-485.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw Hill, 1974.

Peng, C. Y., & Subkoviak, M. J. A note on Huynh's approximation procedure for estimating criterion-referenced reliability. Journal of Educational Measurement, in press.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M. J. Decision-consistency approaches. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Wilcox, R. R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.

DATE

ILMED

-8